Summer 8-2016

Methods to Account for Breed Composition in a Bayesian GWAS Method which Utilizes Haplotype Clusters

Danielle F. Wilson-Wells
University of Nebraska-Lincoln, twinonedil@gmail.com

METHODS TO ACCOUNT FOR BREED COMPOSITION IN A BAYESIAN GWAS METHOD WHICH UTILIZES HAPLOTYPE CLUSTERS

by

Danielle Faye Wilson-Wells

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Statistics

Under the Supervision of Professor Stephen D. Kachman

Lincoln, Nebraska

August, 2016

METHODS TO ACCOUNT FOR BREED COMPOSITION IN A BAYESIAN GWAS METHOD WHICH UTILIZES HAPLOTYPE CLUSTERS

Danielle Faye Wilson-Wells, Ph.D.

University of Nebraska, 2016

Adviser: Stephen D. Kachman

In livestock, prediction of an animal's genetic merit using genomic information is becoming increasingly common. The models used to make these predictions typically assume that we are sampling from a homogeneous population. However, in both commercial and experimental populations the sire and dam of an individual may be a mixture of different breeds. Haplotype models can capture this population structure.

Two models based on breed specific haplotype clusters where developed to account for differences across multiple breeds. The first model utilizes the breed composition of the individual, while the second utilizes the breed composition from the sire and dam. Haplotype clusters were modeled as hidden states in a hidden Markov model where the genomic effects are associated with loci located on the unobserved clusters. Similar to the Bayes C model, we can model the genomic effects at the loci using a prior, which consists of a mixture of a multivariate normal and a point mass at zero distribution.

The performance of the first model will be evaluated in a composite beef cattle population, representing various fractions of several breeds, using five weight traits, seven carcass traits, and two other traits related to calving on 6,552 cattle genotyped for 99,827 mapped SNPs. The performance of the second model will be evaluated in a two-way cross population, which was a

cross between two independent lines, using age of puberty records on 1,654 swine genotyped for 48,408 mapped SNPs. Both models will also be evaluated in a simulated composite population of two lines of 12,500 individuals and 61,255 mapped SNPs.

Overall, the breed specific haplotype models led to larger and more clearly observed estimated QTL. However, the prediction accuracy for the haplotype models were typically lower than those for the traditional Bayesian GWAS models. Therefore, while our ability to locate QTLs was increased, the traditional models are still the preferred choice for prediction as they have higher prediction accuracy when it comes to estimating an animal's genetic merit.

ACKNOWLEDGEMENTS

I would like to first thank my adviser, Dr. Stephen Kachman for all of your patients, wisdom, and encouragement. You pushed me to continue when all I wanted to do was give up. You were always understanding and gave me just the right amount of tough love when it was needed.

I would also like to thank the other members of my committee: Dr. Kent Eskridge, Dr. Qi Zhang, and Dr. Daniel Ciobanu. I want to thank you for your roll in ensuring I completed my dissertation. For some that role was serving on my reading committee. Others it was your lighthearted joking that made my time here at UNL truly unforgettable.

I would also like to thank all of my office mates, past and present. Specifically, I want to thank Elizabeth, Marina, Julie, and Jana. Thank you for always being a listening ear or helping me work through a problem. I can honestly say I would have never made it through my classes, let alone finished my dissertation without you.

To my husband, James Wells, I want to say thank you. You never let me quit and even when I was completely stressed and felt like this was the biggest mistake of my life you stood beside me. You encouraged me, you pushed me, and most of all you supported me. Without you by my side this dissertation would have never been possible.

To my children, Rosabella and Corbin, thank you for being there when mommy needed a snuggle. To my parents, thank you for encouraging me in every endeavor. You made me feel like I could do anything. Finally, to the rest of my family, my siblings and my in-laws thank you for being there for me and encouraging me. Thanks to all you, I was able to something I never even considered before I got to UNL.

GRANT INFORMATION

Research for this dissertation funded by: USDA NIFA grant 2013-68004-20370 and an UNL-USMARC internal grant

PREFACE

Portions of Chapter 3 and Chapter 4 were submitted for publication in the Proceedings for the 28th Annual Conference on Applied Statistics in Agriculture. (D.F. Wilson-Wells and S.D. Kachman, *A Bayesian GWAS Method Utilizing Haplotype Clusters for a Composite Breed Population*, Proceedings for the 28th Annual Conference on Applied Statistics in Agriculture, 2016.)

All figures were generated using R 3.2 [47].

All analyses for Bayes B and Bayes C were preformed using the GenSel Software v4.73 [13].

Table of Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

We are interested in extending a class of Bayesian models based on haplotype clusters. The current class of models are built to predict an individual's genetic merit under the assumption of sampling individuals from a single homogeneous outbred population. Three extensions on the existing models will be developed to allow for a population composed of a mixture of several distinct sub-populations. Current models often perform poorly when dealing with a mixture of sub-populations due to changes in linkage disequilibrium between markers and quantitative trait loci across sub-populations. It is hoped that by using a model based on sub-population specific haplotype clusters the model will do a better job of capturing changes in linkage disequilibrium across sub-populations.

The first extension simply re-weights the haplotype cluster membership probabilities based on the individuals unique breed composition. This is the simplest version and uses a single parameter to control how often a transition occurs between haplotype clusters. The second extension is to include a parameter to control how often a transition occurs between haplotype clusters of the same sub-population and haplotype clusters from different sub-populations. This extension, like the first, utilizes the individuals breed composition in order to re-weight the haplotype cluster membership probabilities. The third

extension is to utilize the parental breed composition information for each individual rather than the individuals breed compositions itself. Similar to the second extension, this extension will include a parameter to control how often a transition occurs between haplotype clusters of the same sub-population and haplotype clusters from different sub-populations. It is hoped that by utilizing the parental breed composition of an individual we will be better able to control the cluster membership probabilities.

CHAPTER 2

THE HISTORY OF GENOMIC SELECTION MODELS

2.1 Introduction

From an animal breeding perspective, individuals are evaluated based on their potential for economic gain. We want to find dairy cattle that produce the most milk, swine that produce the most offspring, or beef cattle that have the highest quality of meat. These traits are a few examples of phenotypic traits that we could select for. Phenotypes can be measured quantitatively or qualitatively and are characteristics of an individual that can be visually observed. Variability in phenotypes have both environmental and genetic sources.

Selective breeding has been utilized for thousands of years, even before the role of genetics was understood. With the advancement of our ability to quantify genetic information, the intensity with which selection is preformed has increased. Selection in large populations of livestock and over many generations has had favorable effects on pushing phenotypic traits in a beneficial direction. One of the goals of genetic research is to be able to map the genetic mechanisms that are controlling the phenotypic variation in order to improve our success in selection [2]. Gregor Mendel in the 1800s laid the framework of modern genetics, but his research was centered toward understanding qualitative traits. Ronald A. Fisher in 1918 was the first to lay the framework

that allows us to account for the variation in theses phenotypes using genetic mapping and genetic analysis of quantitative traits [60]. Genetic evaluation is centered on the analysis of phenotypic and pedigree information to predict the genetic merit of a particular individual, which is quantified by the individuals breeding value. Leif Andersson defines breeding value as "the genetic merit of an individual estimated using the phenotypic deviation of its offspring from the population mean [2]."

Recently, a new source of genomic information in the form of DNA-based markers has become available. One of the first DNA-based markers used were microsatellites which are short, tandem sequences that repeat. Microsatellites are highly abundant in the mammalian genome and highly polymorphic which makes them good candidates for explaining the variance in the phenotype [57]. Another source of genomic information is a Single Nucleotide Polymorphism (SNP) which we believe to be responsible for the variation observed in the phenotype. We can use a single SNP or a SNP panel which consists a large number of SNPs, somewhere between 5,000 and 500,000 [51]. The problem with using only one SNP is that most phenotypic traits are not caused by a single SNP and, therefore, SNP panels are usually preferred. Over time, more and more SNPs have been discovered leading to a larger set of markers that can be used to explain the phenotypic variation. We now have the ability to obtain the complete sequence instead of specific points in the genome. The major disadvantage is that we also receive information from markers that are not polymorphic and have no effect on the phenotypic variation [57].

SNP panels and sequences lead to models where the number of markers considerably outnumber the number of individuals. One of the goals is to identify the markers that are most informative when making predictions. Es-

sentially, which genetic variants give us the best prediction of the breeding value using data that is only collected from hundreds to tens of thousands of individuals [20].

Many models have been developed to perform genetic evaluations, and mixed models have been and still are the underlying fundamental approach to genetic evaluation. Traditionally, these models have fallen into one of two schools of thought. The first are single-step models that utilize the mixed model equations. The second are multi-step models based on a Bayesian framework. The first major difference between these two models is the covariance matrix used. With the single step models, we use a non-diagonal covariance matrix on the breeding values with all covariates included. The multi-step models use a covariance matrix for the markers which is diagonal but allows for a mixture model. The second major difference is how the models handle estimation and prediction. The single-step models allow individuals to remain ungenotyped and performs estimation and prediction in a single-step. The multi-step models require that we do estimation and prediction in two separate steps. First, we estimate the marker effects using only the genotyped individuals. Second, we blend the marker estimates with the phenotypic information on all individuals in order to make predictions. Recently, a third school of thought has developed merging the two existing models together, called single-step Bayes (SS-Bayes). Like a single-step model, SS-Bayes allows individuals with missing genotypes to remain in the analysis and preforms estimation and prediction in a single step. Like a multi-step model, SS-Bayes uses the covariance matrix on the markers themselves which is diagonal and allows for a mixture model.

All three of the above schools begin with the same linear mixed model

equation. In general, the models we will be considering have the form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

or equivalently:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{M}\boldsymbol{\alpha} + \mathbf{e}.$$

We define $\mathbf{y}$ as a $n \times 1$ vector of phenotypes taken from n individuals and $\mathbf{X}$ as an $n \times p$ incidence matrix relating the $p \times 1$ vector $\boldsymbol{\beta}$ of p fixed effects to the individual. Further, $\mathbf{Z}$ is an $n \times n$ incidence matrix relating the individual to its genetic information and $\mathbf{u}$ is a $n \times 1$ vector of breeding values for the n individuals, where $\mathbf{u} = \mathbf{M}\boldsymbol{\alpha}$. We the define $\mathbf{M}$ as an $n \times k$ matrix of genotype covariates for the k SNP markers and $\boldsymbol{\alpha}$ as a $k \times 1$ vector of random regression coefficients of the k SNPs. The k SNPs in the $\mathbf{M}$ matrix are usually coded as 0, 1, or 2 to represents the number of A alleles. Finally, $\mathbf{e}$ is an $n \times 1$ vector of residuals [14].

2.2 Single-Step models

2.2.1 The basic approach

The single-step models fall under the frequentist school of thought in statistics which was developed by Karl Pearson and Ronald Fisher during the 1930s. Other influential statisticians from the frequentist school include Jerzy Neyman and Abraham Wald [4]. The first national implementation of the single-step models was performed in 2010 by Aguilar et al. [1] and utilized U.S. Holstein data from 1955 to 2009.

The single-step models utilize mixed model theory to estimate each markers

effect. We start by considering the model equation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}.$$

Here we use $\mathbf{Z}$ to account for individuals with repeated records or individuals without records and $\mathbf{u} = \mathbf{M}\boldsymbol{\alpha}$. We further assume the following:

$$var(\mathbf{u}) = \mathbf{A}\sigma^2_{\mathrm{u}} \text{ and } var(\boldsymbol{\alpha}) = \mathbf{I}\sigma^2_{\alpha},$$

where $\mathbf{A}$ is an additive relationship matrix based on the pedigree of the individuals which have all been genotyped and $\mathbf{R}$ is a diagonal covariance matrix. We can also define the genomic-based relationship matrix, $\mathbf{G}$, which is formed as the cross-product of the $\mathbf{M}$ matrix scaled by:

$$k = 2\sum p_j \left(1 - p_j\right),$$

where p_j is the allele frequency for marker j which is assumed to be biallelic. Dividing by k puts $\mathbf{G}$ on the same scale as $\mathbf{A}$. That is:

$$\mathbf{G} = \frac{\mathbf{M}\mathbf{M}'}{k}.$$

The matrix $\mathbf{M}$ is centered at zero with elements corresponding to individual i and marker j:

$$m_{ij} = \begin{cases} 0 - 2p_j & \textit{when there are no A alleles} \\ 1 - 2p_j & \textit{when there is one A allele} \\ 2 - 2p_j & \textit{when there are two A alleles} \end{cases}$$

[64]. Finally, we assume that

$$var(\mathbf{e}) = \mathbf{R}\sigma_{\mathrm{e}}^2$$

[39].

In order to get solutions for $\boldsymbol{\beta}$ and $\mathbf{u}$, we can just simply solve the model equation, which produces:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \lambda_u\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix},$$

where $\lambda_u = \frac{\sigma_e^2}{\sigma_u^2}$ [32]. Here we are using the additive relationship matrix, $\mathbf{A}$, if we do not know the additive relationship matrix we can substitute the genomic relationship matrix, $\mathbf{G}$, in place of $\mathbf{A}$ which leads to an alternative form of the mixed model equations, which is:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \lambda_u\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

[64]. If we do not have genotypic information on any of the individuals, but we have pedigree information, we can estimate the individuals breeding values using the first set of equations. If we have genotypic information on all the individuals, but we do not have pedigree information, then we can estimate the individuals breeding values using the second set of equations. A third equivalent set of mixed model equations are:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{M} \\ \mathbf{M}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{M}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{M} + \lambda_\alpha\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{M}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix},$$

where $\lambda_\alpha = \frac{\sigma_e^2}{\sigma_\alpha^2}$. Similar to the second set of equations, the above set requires that we have the genotypic information on all the individuals, but we do not need the pedigree information.

2.2.2 Accounting for Ungenotyped individuals

A major problem with the basic approach above is that we rarely have access to data that includes genotypic information for every individual. Traditionally, the data includes genotypic information on the parents, but there is no available genotypic information for the offspring or vice versa. To get solutions when some of the individuals are not genotyped requires the use of multiple-step procedures, which are known to have bias and errors. Therefore, a modified matrix that accounts for not only the pedigree-based relationships but also accounts for genomic-based relationships and can be computed in a single-step was suggested [39].

First, we partition the model based on whether an individual has or has not been genotyped. Using the subscript 1 to denote the partition containing individuals that have not been genotyped and the subscript 2 to denote the partition containing individuals that have been genotyped. The model becomes:

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_2} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \mathbf{e},$$

where $\mathbf{u_2} = \mathbf{M_2}\boldsymbol{\alpha}$ and $\mathbf{M_2}$ is the matrix of genotype covariates for the k SNP markers for the genotyped individuals. Next, we can write A as:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{bmatrix} \text{ with inverse } \mathbf{A^{-1}} = \begin{bmatrix} \mathbf{A^{11}} & \mathbf{A^{12}} \\ \mathbf{A^{21}} & \mathbf{A^{22}} \end{bmatrix},$$

where $\mathbf{A_{11}}$ is the relationship matrix between all ungenotyped individuals, $\mathbf{A_{22}}$ is the relationship matrix between all genotyped individuals, and $\mathbf{A_{12}}$ and $\mathbf{A_{21}}$ are the relationship matrices between the ungenotyped and genotyped individuals. Using $\mathbf{A^{-1}A = I}$, we can establish several properties of partitioned matrices:

$$\mathbf{A^{11}A_{11} + A^{12}A_{21} = I}, \tag{2.2.1}$$

$$\mathbf{A^{21}A_{12} + A^{22}A_{22} = I}, \tag{2.2.2}$$

$$\mathbf{A^{11}A_{12} + A^{12}A_{22} = 0}, \tag{2.2.3}$$

$$\mathbf{A^{21}A_{11} + A^{22}A_{21} = 0}, \text{ and} \tag{2.2.4}$$

$$\left(\mathbf{A_{11} - A_{12}\left(A_{22}\right)^{-1}A_{21}}\right)^{-1} = \mathbf{A^{11}} \tag{2.2.5}$$

[58]. Finally, we consider the partitioned genomic-based relationship matrix, $\mathbf{G_{22}}$:

$$\mathbf{G_{22}} = \frac{\mathbf{M_2 M_2'}}{k}$$

[1].

Now, using a method develop by Legarra [34], consider $\mathbf{u}_1$ conditioned on the observed genotypes, $\mathbf{M}_2$. We can break $\mathbf{u_1}|\mathbf{M_2}$ into two parts. The first is the expected value of $\mathbf{u_1}$ conditioned on $\mathbf{u_2}$ and the second is an uncorrelated residual, $\boldsymbol{\varepsilon}$. The vector $\mathbf{u}_1$ can be written as follows:

$$\mathbf{u_1}|\mathbf{M_2} = \mathrm{E}\left(\mathbf{u_1}|\mathbf{u_2}\right) + \boldsymbol{\varepsilon} = \mathrm{BLUP}\left(\mathbf{u_1}|\mathbf{u_2}\right) + \boldsymbol{\varepsilon} = \mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-1}\mathbf{M_2}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

since $\mathbf{u}_2 = \mathbf{M}_2\boldsymbol{\alpha}$ as stated above and the distribution $\mathbf{u_1}$ conditioned on $\mathbf{u}_2$ is:

$$\mathbf{u}_1|\mathbf{u_2} \dot{\sim} \mathrm{N}\left(\mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-1}\mathbf{u_2},\ \left(\mathbf{A_{11} - A_{12}}\left(\mathbf{A_{22}}\right)^{-1}\mathbf{A_{21}}\right)\sigma^2_{\mathrm{u}}\right).$$

When we assume normality, then $\mathrm{E}\left(\mathbf{u_1}|\mathbf{u_2}\right)$ is $\mathrm{BLUP}\left(\mathbf{u_1}|\mathbf{u_2}\right)$.

Next, using the fact that:

$$Var\left(\mathbf{u_2}|\mathbf{M}_2\right) = \mathbf{M}_2\mathbf{M}_2'\sigma_\alpha^2 = \frac{\mathbf{M}_2\mathbf{M}_2'}{k}\sigma_u^2 = \mathbf{G}_{22}\sigma_u^2 \text{ and}$$

$$Var\left(\boldsymbol{\varepsilon}\right) = \mathbf{A_{11}} - \mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{A_{21}}\sigma_u^2,$$

the variance of $\mathbf{u}_1$ is:

$$Var\left(\mathbf{u_1}|\mathbf{M}_2\right) = Var\left(\mathbf{u_1}|\mathbf{u_2}\right) + Var(\boldsymbol{\varepsilon})$$

$$= \left(\mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{G_{22}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{A_{21}} + \mathbf{A_{11}} - \mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{A_{21}}\right)\sigma_u^2,$$

since $\boldsymbol{\varepsilon}$ is uncorrelated. Finally, the covariance between $\mathbf{u}_1$ and $\mathbf{u}_2$ is:

$$Cov\left(\mathbf{u_1}, \mathbf{u_2}|\mathbf{M}_2\right) = \mathbf{A}_{12}\left(\mathbf{A}_{22}\right)^{-\mathbf{1}}\mathbf{M}_2\mathbf{M}_2'\sigma_\alpha^2 = \mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{G_{22}}\sigma_u^2.$$

The pedigree-genomic based relationship matrix, called $\mathbf{H}$, we want is formed as follows:

$$\mathbf{H} = Var(\mathbf{u}|\mathbf{M}_2) = \begin{bmatrix} \mathrm{Var}(u_1|\mathbf{M}_2) & \mathrm{Cov}(u_1, u_2|\mathbf{M}_2) \\ \mathrm{Cov}(u_2, u_1|\mathbf{M}_2) & \mathrm{Var}(u_2|\mathbf{M}_2) \end{bmatrix}$$

$$\mathbf{H} = \mathbf{A}\sigma_u^2$$

$$+ \begin{bmatrix} \mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\left(\mathbf{G_{22}} - \mathbf{A_{22}}\right)\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{A_{21}} & \mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\left(\mathbf{G_{22}} - \mathbf{A_{22}}\right) \\ \left(\mathbf{G_{22}} - \mathbf{A_{22}}\right)\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{A_{21}} & \mathbf{G_{22}} - \mathbf{A_{22}} \end{bmatrix}\sigma_u^2$$

[34].

If we take the inverse of $\mathbf{H}$,we find

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{G_{22}})^{-1} - (\mathbf{A_{22}})^{-1} \end{bmatrix}$$

[39]. Then using properties to partitioned matrices established above, Equations 2.2.1 through 2.2.5, it can be shown that $\mathbf{H^{-1}H} = \mathbf{I}$. Once we have formed $\mathbf{H}^{-1}$, the mixed model equations are formed by replacing $\mathbf{A}$ in the original set of equations with $\mathbf{H}$ as follows:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} + \lambda_{\mathrm{u}}\mathbf{H^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{bmatrix}$$

where λ_u is defined above [58].

2.2.3 Alternative Mixed Model Equations

It should be noted that calculating $\mathbf{H^{-1}}$ is dependent on $\mathbf{G_{22}^{-1}}$ existing. When $\mathbf{G_{22}^{-1}}$ does not exist, an alternative set of mixed model equations was suggested by Henderson. This set of equations can be defined as follows:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{HZ'R^{-1}X} & \mathbf{HZ'R^{-1}Z} + \lambda_{\mathrm{u}}\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{HZ'R^{-1}y} \end{bmatrix}$$

[24]. The advantage here is that these equations only require the use of $\mathbf{H}$, which does not have to be full rank and the invertability of $\mathbf{H}$ is no longer an issue.

2.3 Multi-step (Bayesian) models

2.3.1 The Bayesian Framework

The current school of Bayesian statistics began when Count Laplace published several articles related to the matter between 1774 and 1812. However, Thomas Bayes introduced the same ideas several years before Laplace, which is why he is credited with its discovery. Some argue that the true founder of this approach is Saunderson, who is responsible for publishing the principle that Bayesian inference is based on before either Bayes or Laplace [4]. Bayesian inference models were introduced into an individual breeding framework by Daniel Gianola and J.L. Foulley in 1982 [17].

The objective of Bayesian inference is to combine our prior information with the observed data to quantify the uncertainty, measured with a distribution, about the true value for a parameter [4]. Thus Bayesian inference is based on finding the distribution of the parameters given the data. We will let $\boldsymbol{\theta}$ be the vector of unknown parameters and $\mathbf{y}$ the vector of observations. Let $f(\boldsymbol{\theta}|\mathbf{y})$ be the posterior distribution of the parameters given the data, $L(\boldsymbol{\theta}|\mathbf{y})$ the likelihood function of the parameters, and $f(\boldsymbol{\theta})$ the prior distribution of the parameters. Using Bayes theorem, the posterior distribution can be written in terms of the likelihood and the prior distribution of the parameters, where $f(\mathbf{y})$ is the marginal distribution of $\mathbf{y}$ and is equal to:

$$f(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

Then the posterior distribution is:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})}{f(\mathbf{y})} \propto f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{y}) \cdot f(\boldsymbol{\theta})$$

[4].

One crucial aspect of this procedure is specifying the likelihood function and prior distributions. In the case of genomic prediction, we have to assign a likelihood function for our phenotypic trait data as well as prior distributions for our marker effects and various variance components that need to be estimated. Using the assigned prior and likelihood function produces the posterior distribution for the unknown genetic merit of an individual. The posterior distribution combines information about the genetic structure for the phenotypic trait of interest with the observed phenotypic and genomic data [67]. The approaches that follow are Bayesian regression models. The main goal is to select the markers which best predict an individual's genetic merit for the phenotypic trait of interest based on a combination of phenotypic and genomic information.

We will consider the five most widely used Bayesian models for whole genome prediction. The following Bayesian regression models share a common model equation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZM}\boldsymbol{\alpha} + \mathbf{e},$$

which is defined above and where $\mathbf{Z}$ is an identity matrix. In addition to the model equation above, there are several similarities between all of the models. First, all models place an improper flat prior on $\boldsymbol{\beta}$, which is denoted as:

$$\boldsymbol{\beta} \sim U(-\infty, \infty).$$

Second, we assume that the random effects, $\mathbf{e}$, have a multivariate normal prior distribution with a mean of $\mathbf{0}$ and a covariance matrix of $\mathbf{R}\sigma_e^2$, where $\mathbf{R}$ is a diagonal matrix. We will denote this as:

$$\mathbf{e}|\mathbf{R}\sigma_e^2 \sim N(\mathbf{0}, \mathbf{R}\sigma_e^2)$$

[14]. Third, for σ_e^2 we use a scaled inverse chi-squared prior distribution with known scale parameter, s_e^2, and known degrees of freedom, ν_e, which is denoted as:

$$\sigma_e^2|\nu_e, s_e^2 \sim Inv - \chi^2(\nu_e, s_e^2)$$

[38]. In general, a scaled inverse chi-squared distribution with degrees of freedom ν and scale parameter s^2 has density function:

$$p(\theta) = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma\left(\frac{\nu}{2}\right)} s^{\nu} \theta^{-\left(\frac{\nu}{2}+1\right)} e^{-\frac{\nu s^2}{2\theta}},$$

with expected value:

$$E(\theta) = \frac{\nu}{\nu - 2} s^2.$$

Finally, the likelihood for the phenotypic trait, where $\boldsymbol{\theta}$ is the vector of unknown fixed and random effects, is assumed to have a multivariate normal distribution with a mean of $\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\alpha}$ and a covariance of $\mathbf{R}\sigma_e^2$. We denote this as follows:

$$\mathbf{y}|\boldsymbol{\theta} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\alpha}, \mathbf{R}\sigma_e^2)$$

[14]. Thus, the following models will only differ in their prior distribution for $\boldsymbol{\alpha}$.

2.3.2 Random Regression-Best Linear Unbiased Prediction

Random regression-best linear unbiased prediction (RR-BLUP) is the first model of Bayesian estimation we will discuss. Rønningen in 1971 [48] and Robinson in 1991 [49] each suggested that best linear unbiased prediction (BLUP) can be thought of as a Bayesian estimator. Using a BLUP of the random effects has many benefits. First, BLUP is the standard prediction of genetic merit; therefore, there is a long history of well developed software and familiarity with the algorithms used. Next, BLUP has good optimality properties as it minimizes the prediction error variance among all linear unbiased predictors [40]. Finally, it only requires that we know the first and second moments. However, we must require that the marker effects have a normally distributed prior, since the normality assumption guarantees that the BLUP satisfies all the conditions to be the posterior mean, which is the optimal estimator under squared loss [14].

In RR-BLUP, we assign the fixed effects, $\boldsymbol{\beta}$, to have a flat prior as was mentioned above. Then the marker effects, $\boldsymbol{\alpha}$, are assigned a normally distributed prior which is independent of the fixed effects. Thus, a single marker has a prior that is a normal distribution with a mean of 0 and variance equal to σ^2_α denoted as:

$$\alpha_i|\sigma^2_\alpha \sim N(0, \sigma^2_\alpha).$$

For σ^2_α, we use a scaled inverse chi-squared prior distribution with known scale parameter, s^2_α, and known degrees of freedom, ν_α, which is denoted as:

$$\sigma^2_\alpha|\nu_\alpha, s^2_\alpha \sim Inv - \chi^2(\nu_\alpha, s^2_\alpha) \tag{2.3.1}$$

[14]. When the ratio $\lambda_\alpha = \frac{\sigma_e^2}{\sigma_\alpha^2}$ is known, RR-BLUP produces the BLUP of $\boldsymbol{\alpha}$, and this can easily be obtained by solving the mixed model equations [24]. When λ is unknown independent priors for σ_α^2 and σ_e^2 are added to the model.

2.3.3 Bayes A

The second Bayesian model is Bayes A. The difference between Bayes A and RR-BLUP is that in Bayes A each marker has a unique variance. For Bayes A, we say that a single marker has an effect that is a normal prior distribution with a mean of 0 and variance equal to σ_i^2, denoted as:

$$\alpha_i|\sigma_i^2 \sim N(0, \sigma_i^2).$$

We also assume that the marker effects are independent of each other. Finally, we assume the variance for each marker effect, σ_i^2, has a scaled inverse chi-squared prior distribution with known scale parameter, s_α^2, and known degrees of freedom, ν_α, as shown below:

$$\sigma_i^2|\nu_\alpha, s_\alpha^2 \sim Inv - \chi^2(\nu_\alpha, s_\alpha^2) \tag{2.3.2}$$

[38].

2.3.4 Bayes B

The third model is called Bayes B. Bayes B differs from Bayes A in that instead of assuming all markers have a normal prior, we will assume that a proportion, π, have an effect of 0. We are assigning each marker to a mixture

prior distribution. The marker has a probability of π of having an effect of 0 and a probability of $1-\pi$ of being normally distributed with a mean of 0 and variance of σ_i^2, which we denote as

$$\alpha_i|\sigma_i^2 \sim \begin{cases} 0 & \text{with probability } \pi \\ N(0,\sigma_i^2) & \text{with probability } 1-\pi \end{cases}.$$

Similar to Bayes A, the variance of each effect is defined the same as it was above in 2.3.2. It should also be noted that Bayes A is just a special case of Bayes B when $\pi = 0$ [38].

2.3.5 Bayes C and Bayes Cπ

The next model is called Bayes C. Bayes C differs from Bayes B in that it assumes that the markers are sampled from a normal prior distribution with equal variance. Similar to Bayes B, we will assign each marker a mixture prior distribution. The marker has a probability of π of having an effect of 0 and a probability of $1-\pi$ of being normally distributed with a mean of 0 and variance of σ_α^2, which we denote as

$$\alpha_i|\sigma_\alpha^2 \sim \begin{cases} 0 & \text{with probability } \pi \\ N(0,\sigma_\alpha^2) & \text{with probability } 1-\pi \end{cases}.$$

Then, similar to RR-BLUP, we assume that this variance, σ_α^2, has the same prior that was defined in 2.3.1 [37].

Finally, Bayes Cπ differs from Bayes C by adding a prior for π. Typically,

we assume that π is an unknown quantity with a uniform prior shown below:

$$\pi \sim Unif(0,1)$$

[21]. An alternative prior, suggested by Pérez and de los Campos [46], is to assign π a beta prior where:

$$\pi \sim Beta\,(p_0, \pi_0)$$

with distribution function:

$$p(\pi) = \frac{\Gamma\,(p_0)}{\Gamma\,(p_0\pi_0)\,\Gamma\,(p_0 - p_0\pi_0)}\pi^{(p_0\pi_0-1)}\,(1-\pi)^{(p_0-p_0\pi_0-1)}$$

and:

$$E(\pi) = \pi_0 \text{ and } Var(\pi) = \frac{\pi_0\,(1-\pi_0)}{(p_0+1)}.$$

If we set $\pi_0 = 0.5$ and $p_0 = 2$, then the beta prior reduces down and become the traditional uniform prior. Also, if we let p_0 be large, then the prior reduces to the point estimate π_0.

2.3.6 A Comparison of the Bayesian models

Many papers have been published whose goal was to determine which of the above models produced the most accurate genomic predictions. The conclusion is that there is no model which is uniformly most accurate. Meuwissen et al. [38] used simulated data to demonstrate that Bayes B outperforms both Bayes A and the single step models when the marker density is high. Habier et al. [21] used simulations based on the North American Holstein data to show that

Bayes A outperforms Bayes B, Bayes C, RR-BLUP and the single-step models when there are only a few markers which have a larger effect. Daetwyler et al. [10] used simulated data to demonstrate that when the number of marker which have a large effect is large the single step model gives better predictions than Bayes C. Clark et al. [9] used simulated data to show that Bayes B outperforms the single-step model when different distributions are assigned to the allele frequency.

In Table 2.3.1 we compare the priors used for each of the Bayesian models. From the table it is clear that the prior distribution of the marker effects is going to effect which Bayesian model preforms better. Sample size, number of markers in the model, genetic architecture of the trait and individuals, and many other factors also effect the overall performance of a model. Each model has its strengths and weaknesses. When the data matches the strengths, that particular model will perform better. The most appropriate model is dependent on the underling genetic architecture of the trait itself.

2.4 A Merger of Models

2.4.1 The Motivation

Single-step models and the Bayesian models each have their own set of benefits. A hybrid model that merges the two models can take advantage of both sets of benefits. One such hybrid model is called single-step Bayes (SS-Bayes). SS-Bayes includes both the single-step models and the Bayesian models as special cases. The idea of a hybrid model is not new and has been used for breeding value prediction, gene selection, phenotype prediction, and mapping complex traits. However, SS-Bayes uses an algorithm which efficiently handles

Table 2.3.1: Prior Distributions of the Bayesian models

Model	**Prior for Marker effect**	**Prior for marker variance**	**Prior for error variance**
RR-BLUP	$\alpha_i \vert \sigma^2_\alpha \sim N(0, \sigma^2_\alpha)$	$\sigma^2_\alpha \vert S^2_\alpha \sim \chi^{-2}(\nu_\alpha)$	$\sigma^2_e \vert S^2_e \sim \chi^{-2}(\nu_e)$
Bayes A	$\alpha_i \vert \sigma^2_i \sim N(0, \sigma^2_i)$	$\sigma^2_i \vert S^2_\alpha \sim \chi^{-2}(\nu_\alpha)$	$\sigma^2_e \vert S^2_e \sim \chi^{-2}(\nu_e)$
Bayes B	$\alpha_i \vert \sigma^2_i \sim$ $\begin{cases} 0 & \text{with probability } \pi \\ N(0, \sigma^2_i) & \text{with probability} 1-\pi \end{cases}$	$\sigma^2_i \vert S^2_\alpha \sim \chi^{-2}(\nu_\alpha)$	$\sigma^2_e \vert S^2_e \sim \chi^{-2}(\nu_e)$
Bayes C	$\alpha_i \vert \sigma^2_\alpha \sim$ $\begin{cases} 0 & \text{with probability } \pi \\ N(0, \sigma^2_\alpha) & \text{with probability } 1-\pi \end{cases}$	$\sigma^2_\alpha \vert S^2_\alpha \sim \chi^{-2}(\nu_\alpha)$	$\sigma^2_e \vert S^2_e \sim \chi^{-2}(\nu_e)$
Bayes Cπ	$\alpha_i \vert \sigma^2_\alpha \sim$ $\begin{cases} 0 & \text{with probability } \pi \\ N(0, \sigma^2_\alpha) & \text{with probability } 1-\pi \\ & \pi \sim U(0,1) \end{cases}$	$\sigma^2_\alpha \vert S^2_\alpha \sim \chi^{-2}(\nu_\alpha)$	$\sigma^2_e \vert S^2_e \sim \chi^{-2}(\nu_e)$

data sets with more individuals and more markers better than the previous applications did from a computing perspective [68].

SS-Bayes, like the single-step models, combines phenotype, genotype, and pedigree information together into one step. Like the multi-step models, SS-Bayes does not require that the marker effects be normally distributed [15]. We can use mixture distributions or non-normal distributions. SS-Bayes uses the relationship matrix, $\mathbf{A}^{-1}$, which is preferred since $\mathbf{A}$ is a sparse matrix which is always invertible unless the individuals have clones. In addition, the computational time to compute $\mathbf{A}^{-1}$ increases linearly as the number of individuals increases, whereas the computational time to compute $\mathbf{G}^{-1}$ increases cubically as the number of genotyped individuals increases [15]. A difference between the SS-Bayes model and the above mentioned single-step and Bayesian models is that an additional error term is introduced in order to account for differences between the imputed genotypes and the actual unobserved genotypes [15].

When considering a new model, we want a model that performs better than what we already have and at worst performs just as well as the current model. Zhou [68] compared SS-Bayes to both the single-step models and the Bayesian models. He found that SS-Bayes either performed as well or slightly better than both the single-step models and the Bayes models.

2.4.2 SS-Bayes

2.4.2.1 Building the model equation

Starting with the same model equation used in both the single-step and Bayesian models:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZM}\boldsymbol{\alpha} + \mathbf{e} \text{ or } \mathbf{y}{=}\mathbf{X}\boldsymbol{\beta}{+}\mathbf{Zu}{+}\mathbf{e},$$

where $\mathbf{Z}$ is an $n \times n$ incidence matrix which accommodates individuals with repeated records or individuals without records. Partitioning the records into records representing genotyped and ungenotyped individuals, as in the single step models, yields this set of model equations:

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_2} \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix} + \mathbf{e}.$$

Now, we replace the missing breeding values, $\mathbf{u}_1$, with the imputed breeding values, $\tilde{\mathbf{u}}_1$ plus an error term:

$$\mathbf{u}_1 = \tilde{\mathbf{u}}_1 + \boldsymbol{\varepsilon} = \text{BLUP}\left(\mathbf{u_1}|\mathbf{u_2}\right) + \boldsymbol{\varepsilon} = \mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{M_2}\boldsymbol{\alpha} + \boldsymbol{\varepsilon},$$

where, like the single step-models:

$$\mathbf{u}_1|\mathbf{u_2} \dot{\sim} \text{N}\left(\mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{u_2},\ \left(\mathbf{A_{11}} - \mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{A_{21}}\right)\sigma^{\mathbf{2}}_{\mathbf{u}}\right)$$

and

$$\varepsilon \dot{\sim} N\left(\mathbf{0}, \mathbf{A_{11}} - \mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-\mathbf{1}}\mathbf{A_{21}}\sigma^2_u\right).$$

Then the imputed matrix of genotype covariates,$\tilde{\mathbf{M}}_\mathbf{1}$, is:

$$\tilde{\mathbf{M}}_\mathbf{1} = \mathbf{A_{12}}\left(\mathbf{A_{22}}\right)^{-1}\mathbf{M_2}.$$

Therefore, the model equations can be written as:

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_2} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{M}}_\mathbf{1}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \\ \mathbf{M_2}\boldsymbol{\alpha} \end{bmatrix} + \mathbf{e}$$

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_2} \end{bmatrix} \begin{bmatrix} \mathbf{A_{12}}(\mathbf{A_{22}})^{-\mathbf{1}}\mathbf{M_2}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \\ \mathbf{M_2}\boldsymbol{\alpha} \end{bmatrix} + \mathbf{e}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{U}\boldsymbol{\varepsilon} + \mathbf{e},$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1\tilde{\mathbf{M}}_1 \\ \mathbf{Z}_2\mathbf{M}_2 \end{bmatrix} \text{ and } \mathbf{U} = \begin{bmatrix} \mathbf{Z}_1 \\ 0 \end{bmatrix}$$

[15].

2.4.2.2 The Bayesian Aspect

To define the Bayesian aspect of the SS-Bayes model is just a matter of assigning priors. We need to assign priors to $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\varepsilon}$, and $\mathbf{e}$. As with the multi-step Bayesian models we will assign a flat prior for our fixed effects, $\boldsymbol{\beta}$. The prior for the marker effects, $\boldsymbol{\alpha}$, can either be a normal prior, as in RR-BLUP and Bayes A, or a mixture of a point-mass at zero and a normal, as in Bayes B and Bayes C. But our choice of prior is not limited to the priors defined by the Bayesian models above. We can use non-normal priors on our marker effects if that is the most reasonable prior. For example, another type of Bayesian regression model, Bayesian Lasso, places a double exponential prior on the marker effects [43]. However, regardless of the prior assigned to $\boldsymbol{\alpha}$ the prior for $\boldsymbol{\varepsilon}$ can be approximated by a multivariate normal distribution with mean $\mathbf{0}$ and a covariance matrix of $(\mathbf{A_{11}} - \mathbf{A_{12}}(\mathbf{A_{22}})^{-\mathbf{1}}\mathbf{A_{21}})\sigma_u^2$ and σ_u^2 has a scaled chi-squared prior distribution with known scale parameter, S_u^2, and known degrees of freedom, ν_u [15]. Like the Bayesian models, the error term, $\mathbf{e}$, has a multivariate normal prior with a mean of $\mathbf{0}$ and variance $\mathbf{R}\sigma_e^2$. Finally, σ_e^2 has

a scaled chi-squared prior distribution with known scale parameter, S_e^2, and known degrees of freedom, ν_e.

2.4.2.3 The Single-Step Aspect

When the variance components, σ_u^2, σ_α^2, and σ_e^2 are known we can simply solve the model equation:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}W} & \mathbf{X_1'R^{11}Z_1} \\ \mathbf{W'R^{-1}X} & \mathbf{W'R^{-1}W} + \mathbf{I}\lambda_\alpha & \mathbf{W_1'R^{11}Z_1} \\ \mathbf{Z_1'R^{11}X_1} & \mathbf{Z_1'R^{11}W_1} & \mathbf{Z_1'R^{11}Z_1} + \mathbf{A^{11}}\lambda_u \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\varepsilon}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{W'R^{-1}y} \\ \mathbf{Z_1'R^{11}y_1} \end{bmatrix},$$

where $\lambda_\alpha = \frac{\sigma_e^2}{\sigma_\alpha^2}$ and $\lambda_u = \frac{\sigma_e^2}{\sigma_u^2}$ as before and $\mathbf{R}^{11}$ represents the portion of $\mathbf{R}^{-1}$ corresponding to the ungenotyped individuals. The solution found will be identical to the solution found using the single-step model where the predictions for $\mathbf{u}$ can be obtained as follows:

$$\hat{\mathbf{u}} = \begin{bmatrix} \tilde{\mathbf{M}}_1 \\ \mathbf{M}_2 \end{bmatrix} \hat{\boldsymbol{\alpha}} + \mathbf{U}\hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} \tilde{\mathbf{M}}_1 \\ \mathbf{M}_2 \end{bmatrix} \hat{\boldsymbol{\alpha}} + \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{0} \end{bmatrix} \hat{\boldsymbol{\varepsilon}}$$

[15]. These mixed model equations are also equivalent to Bayes C when $\pi = 0$ and the variances are known.

2.5 Haplotype Models

2.5.1 The Use of Haplotype Models

As technology has advanced, the size of the available SNP panels has increased. Thus, the number of covariates used in the models above have increased, which in turn increases the required computing time. One way to address this issue is by fitting haplotypes constructed from phased SNP genotypes, rather than fitting the SNPs themselves, where a haplotype is a particular combination of successive marker alleles on a chromosome. An additional advantage of haplotypes over SNPs is that, while SNPs typically included on a SNP panel cannot be in high linkage disequilibrium (LD) with a rare quantitative trait loci (QTL), haplotypes can be in high LD with a rare QTL [19]. This is because SNPs are typically only chosen for SNP panels if their minor allele frequency (MAF) is high and, since high LD can only exist between two loci with similar MAF, SNPs which are in high LD with rare QTL are usually excluded from the SNP panel [62]. Thus the use of haplotypes gives us the potential of being able to detect the rare QTLs that were not detected by SNP based markers.

Haplotype models break the genome up into haplotype blocks. These blocks can be either evenly spaced or be formed such that each haplotype block has an equal number of SNPs. As long as the number of SNPs contained within a haplotype block is larger than the number of haplotypes within the block the overall size of the haplotype model is reduced from the SNP version.

Haplotype models still use the same linear mixed model equation we defined above. The difference is in how the breeding values, $\mathbf{u}$, are calculated. Recall, we defined the model equation as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

or equivalently:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{M}\boldsymbol{\alpha} + \mathbf{e}.$$

Before, $\mathbf{u} = \mathbf{M}\boldsymbol{\alpha}$, where $\mathbf{M}$ is an $n \times k$ matrix of genotype covariates for the n individuals and k SNP markers and $\boldsymbol{\alpha}$ as a $k \times 1$ vector of random regression coefficients of the k SNPs. For the haplotype model, $\mathbf{u} = \mathbf{M_H}\boldsymbol{b}$, where $\mathbf{M_H}$ is an $n \times K_H$ matrix of haplotype covariates and $\boldsymbol{b}$ is a K_H vector of random regression coefficients of the K_H haplotypes. Further, n is the number of individuals, $K_H = \sum_{h=1}^{H} k_h$ is the total number of haplotypes, and k_h is the number of haplotypes in block h.

As mentioned above, haplotype models use haplotypes constructed from phased SNP genotypes. A common approach to phasing SNP genotypes is to use a hidden Markov model (HMM) to model the individuals haplotype. A HMM is composed of a Markov chain for a sequence of unobserved, or hidden, states where each state emits an observed value according to a distribution function. The mathematical development of HMMs traces back to Leonard E. Baum and Ted Petrie in 1966 [3]. The forward-backward algorithms, which are used to efficiently compute the probabilities of the hidden states given the observed values, were described by Ruslan L. Stratonovich in 1960 [61]. In a genetics framework, HMMs have been used for haplotype phasing, haplotype prediction, imputation, association studies, and genomic selection [11]. We will focus on the application of HMMs in haplotype phasing and imputation. Since individuals may be genotyped on different sets of markers, imputation allows us to predict the missing genotypes based on a combination of the in-

dividual's observed genotypes and the pattern of genotypes observed in the other individuals. We can also use imputation to generate genotypes for individuals that we have phenotypic information for, but were not genotyped at all assuming we have genetic information on some of their relatives. Haplotype phasing is where we use the genotypic information in order to infer the maternal and paternal haplotype of an individual.

2.5.2 Description of a Markov Chain

As the HMM uses a Markov chain to model the hidden states, we will start by describing a Markov chain. A Markov chain is a sequence of N states which we will denote as $S = \{S_1, S_2, \ldots, S_N\}$. We will label the underlying state path as $\pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$, where π_k represents the state at locus k and n represents the number of markers. A transition probability, $a(k)_{S_l S_m}$, is the probability that the sequence transitions to state S_m at locus k given the sequence was in state S_l at locus $k-1$, denoted as:

$$a(k)_{S_l S_m} = P(\pi_k = S_m | \pi_{k-1} = S_l).$$

The probability that the initial state, π_1, is S_l is defined to be:

$$a(1)_{0 S_l} = P(\pi_1 = S_l).$$

The probability of a given sequence, π, which has n observations, is:

$$P(\pi) = P(\pi_n, \pi_{n-1}, \ldots, \pi_1).$$

Applying Bayes theorem, the probability of π can be written as:

$$P(\pi) = P(\pi_n|\pi_{n-1}, \ldots, \pi_1) P(\pi_{n-1}|\pi_{n-2}, \ldots, \pi_1) \cdots P(\pi_1).$$

Now, a key property of a Markov chain is that it is memory-less, in that the probability of being in a given state at locus k given all previous states is only dependent on the state at locus $k-1$. Thus the probability of a given sequence, π, is:

$$P(\pi) = P(\pi_n|\pi_{n-1}) P(\pi_{n-1}|\pi_{n-2}) \cdots P(\pi_2|\pi_1) P(\pi_1)$$

$$= P(\pi_1) \prod_{k=2}^{n} P\left(\pi_k|\pi_{(k-1)}\right) = a(1)_{0\pi_1} \prod_{k=2}^{n} a(k)_{\pi_{(k-1)}\pi_k}$$

[12].

2.5.3 Hidden Markov Models

In a hidden Markov model, the states are not observed, instead we observe values emitted by the states. A HMM includes a model for the observed sequence given the hidden states. We will label the q observed emitted values as $O = \{o_1, o_2, \ldots, o_q\}$ and the observed sequence as $X = \{x_1, x_2, \ldots, x_n\}$, where x_k represents the value emitted at locus k. The probability that the sequence emits a particular value at locus k given the hidden state at locus k is called an emission probability. The probability that symbol o_j is seen when in state S_l at locus k is denoted:

$$e_{S_l}(o_j) = P(x_k = o_j|\pi_k = S_l).$$

Similar to the model for the hidden states, the probability of observing value o_i at locus k given the hidden states and previous values up to locus k only depends on the hidden state at locus k. The joint probability of observing the sequence of values, X, and the sequence of states, π is:

$$P(X,\pi) = P(x_n,\ldots,x_1,\pi_n,\ldots,\pi_1)$$

$$= P(x_n|x_{n-1},\ldots,x_1,\pi_n,\ldots,\pi_1)\cdots P(x_1|\pi_n,\ldots,\pi_1)\,P(\pi_n|\pi_{n-1},\ldots,\pi_1)\cdots P(\pi_1)$$

$$= P(x_n|\pi_n)\cdots P(x_1|\pi_1)\,P(\pi_n|\pi_{n-1})\cdots P(\pi_1)$$

$$= a(1)_{0\pi_1}\prod_{k=1}^{n} e_{\pi_k}(x_k)\,a(k)_{\pi_k\pi_{(k+1)}},$$

where $\pi_{n+1} = 0$ [12].

2.5.4 Haplotype Phasing and Imputation Procedures

As was mentioned above, imputation is essential in that it allows us to perform genomic evaluations on individuals which were only genotyped on a subset of the possible markers. The advantage is that we lower costs by genotyping fewer markers, but are still able to obtain information on a larger set. On the other hand, haplotype phasing allows us to reduce our number of overall covariates by breaking the SNPs up into haplotype blocks and using the unique haplotypes as covariate. Current methods can be divided into methods based on family information, methods based on population information, and methods which use both family and population information. Family based information methods are appropriate when individuals have genotyped relatives since we are utilizing linkage and Mendelian segregation rules. Population based infor-

mation methods are appropriate when we have a set of unrelated individuals or individuals whose close relatives are not genotyped since population based imputation utilizes linkage disequilibrium between missing markers and the nearby observed markers [27].

One method available for use is fastPHASE [56]. FastPHASE is a population based method based on a hidden Markov model which performs local clustering of haplotypes at each marker locus. In fastPHASE, the number of clusters is fixed with each cluster representing a hidden state in the HMM [23]. Other programs such as BEAGLE, IMPUTE, and MACH also utilize HMMs; however, we will focus on the method used by fastPHASE.

FastPHASE uses a set of N states which are represented by our N clusters made up of closely related haplotypes. We will denote these clusters as $S = \{S_1, \ldots . S_N\}$. We will be using genotypic data on n loci and P individuals, thus our observed sequence of values for individual i is $X_i = \{x_{i1}, \ldots, x_{in}\}$, where x_{ik} represents the genotype at locus k for individual i. The genotype is equal to the sum of its alleles, thus x_{ik} will have emitted values $O = \{0, 1, 2, .\}$, where $\{.\}$ represents a missing genotype value. Further, each genotype is made up of two haplotypes, the paternal and maternal haplotype. Thus, we can think of each genotype as being emitted from two haplotype clusters. Let π_{ik} denote the unordered pair of clusters with which genotype x_{ik} originates. Therefore, we can think of $\pi_i = \{\pi_{i1}, \ldots, \pi_{in}\}$ as the state path for individual i. As with any HMM, the HMM is defined by the initial state probabilities, transition probabilities, and emission probabilities which we will now define.

The initial-state probability, $a(1)_{0\pi_{i1}}$, is the probability that individual i has the haplotype cluster pair $\pi_{i1} = \{S_l, S_m\}$ at locus 1. The initial state

probability can be written:

$$a\left(1\right)_{0\pi_{i1}} = P_1\left(\pi_{i1} = \{S_l, S_m\}\right) = \begin{cases} \left(\alpha_{S_l 1}\right)^2 & S_l = S_m \\ 2\alpha_{S_l 1}\alpha_{S_m 1} & S_l \neq S_m \end{cases},$$

where $\alpha_{S_l 1}$ is the relative frequency of cluster S_l at locus 1.

The transition probabilities between cluster pairs are constructed from the transition probabilities between the individual clusters within the cluster pairs. We define the probability that individual i transitions from cluster S_l at locus $k-1$ to cluster S_m at locus k on a chromosome as follows:

$$a\left(k\right)_{S_l S_m} = P_k\left(S_l \to S_m\right) = P_k\left(S_m \in \pi_{ik} | S_l \in \pi_{i(k-1)}, \alpha, r\right)$$

$$= \begin{cases} e^{-r_k d_k} + \left(1 - e^{-r_k d_k}\right)\alpha_{S_m k} & S_l = S_m \\ \left(1 - e^{-r_k d_k}\right)\alpha_{S_m k} & S_l \neq S_m \end{cases},$$

where $\alpha_{S_m k}$ is the relative frequency of cluster S_m at locus k, d_k is the physical distance between locus $k-1$ and k, and where $r = (r_2, \ldots, r_n)$ and $\alpha = (\alpha_{S_m k})$ are parameters which need to be estimated. The cluster pair transition probability, $a\left(k\right)_{\pi_{i(k-1)}\pi_{ik}}$, is the probability that individual i transitions from the cluster pair $\pi_{i(k-1)} = \{S_l, S_m\}$ at locus $k-1$ to the cluster pair $\pi_{ik} = \{S_{l'}, S_{m'}\}$ at locus k on the chromosome and can be written:

$$
\begin{aligned}
a(k)_{\pi_{i(k-1)}\pi_{ik}} &= P_k\left(\{S_l, S_m\} \rightarrow \{S_{l'}, S_{m'}\}\right) \\
&= P_k\left(\pi_{ik} = \{S_{l'}, S_{m'}\} \,|\pi_{i(k-1)} = \{S_l, S_m\}\right) \\
&= \begin{cases} a(k)_{S_l S_{l'}}\, a(k)_{S_m S_{m'}} + a(k)_{S_l S_{m'}}\, a(k)_{S_m S_{l'}} & S_l \neq S_m \text{ and } S_{l'} \neq S_{m'} \\ a(k)_{S_l S_{l'}}\, a(k)_{S_m S_{m'}} & \text{otherwise} \end{cases}
\end{aligned}
$$

We can also establish the emission probabilities. The emission probability, $e_{\pi_{ik}}(x_{ik})$, is the probability that individual i at locus k has observed genotype x_{ik} given we are in cluster pair $\pi_{ik} = \{S_l, S_m\}$. The emission probability is written:

$$
\begin{aligned}
e_{\pi_{ik}}(x_{ik}) &= P_k\left(x_{ik} | \pi_{ik} = \{S_l, S_m\}, \theta\right) \\
&= \begin{cases} (1-\theta_{S_l k})(1-\theta_{S_m k}) & x_{ik} = 0 \\ \theta_{S_l k}(1-\theta_{S_m k}) + \theta_{S_m k}(1-\theta_{S_l k}) & x_{ik} = 1 \\ \theta_{S_l k}\theta_{S_m k} & x_{ik} = 2 \end{cases},
\end{aligned}
$$

where $\theta_{S_l k}$ is the allele frequency of the A allele in cluster S_l at locus k [56].

Accuracy of imputed genotypes is crucial to evaluating any imputation method. Marchini and Howie [36] compared fastPHASE to BEAGLE and IMPUTE. They used all three methods to impute 22,270 SNPs that were present in some samples but missing in others. They found that BEAGLE had an error rate of 6.33%, fastPHASE had an error rate of 5.92%, and IMPUTE had an error rate of 5.16%. Pei et al. [45] showed that BEAGLE, IMPUTE, and fastPHASE had error rates that were very similar. They also demonstrated that the higher the linkage disequilibrium, the lower the error rate. Additionally, the showed the higher the density of the markers, the lower the error

rate. Hayes et al. [23] performed imputation on several different breeds of sheep. They showed that the error rate for fastPHASE was lower when imputation was performed within a breed than when it was performed using a combined breed reference population. Weigel et al. [65] compared fastPHASE to IMPUTE and showed that as the number to markers which need imputing decreases, so does the error rate. They also showed that IMPUTE has a lower error rate when the proportion of markers needing imputed is high and that fastPHASE has a lower error rate as long at fewer than 80% of the markers need imputed.

2.5.5 Bayes IM

Current models of Bayesian genome-wide association studies (GWAS) utilize SNPs to find potential QTLs. A true QTL may be located at a locus with which we have not genotyped an individual. The identification of a QTL is dependent on the location of the SNPs as well as their association with the QTL. Ideally, we want to be able to identify these QTL based on the pattern of the SNPs that we have located around the putative QTL alleles.

Kachman [28] proposed a model that uses the information from the SNPs close to the putative QTL which he calls Bayes IM. Bayes IM uses the genotypes from the SNP data and then evenly spaced putative QTL are added along the chromosome, which form haplotype blocks. These putative QTL can be placed at a SNP location or between SNP locations. Next we can use a hidden Markov model similar to the one implemented by fastPHASE [56]. Using a set of haplotype clusters, we can estimate the probability of cluster membership at a locus. Using the parameter estimates obtained from the

hidden Markov model, we can model the genomic effects at the loci using a prior which consists of a mixture of a multivariate normal and a point mass at zero distribution, similar to the Bayes C model. A complete description of the model will be given in the Section 2.5.5.1 and 2.5.5.2.

2.5.5.1 Construct the Haplotype Model

We start by considering the haplotypes that make up an individuals genotype. Each individuals genotype is made up of two haplotypes, one from the mother and one from the father. These haplotypes themselves come from a population of haplotypes, where some are more common than others and some may be very rare. In order to incorporate the haplotypes into the model, we partition the genome into segments, identify the haplotypes within these segments and use the identified haplotypes as covariates.

One option is to treat each possible haplotype as unique. However, unless each segment is short, there will be many unique haplotypes within a segment and some of the possible haplotypes may not be observed within our samples. A second option is to cluster the haplotypes together based on similarity and characterize each cluster based on the frequency of the A allele at each locus. The advantage of this option is that we have fewer covariates.

Next, we shift the segment down by one locus. Using the principle of crossing over, we expect to be able to identify recombination locations as we move down the chromosome. For an original segment and its overlapping shifted segment, the number of common haplotypes should be similar. Based on the common haplotypes in the original segment, we can predict what all but the last locus will be in the common haplotypes of the shifted segment.

We can then create a continuous haplotype model by extending the hap-

lotype clusters across each chromosome. For a particular individual, we break the paternal and maternal haplotypes up into segments with each segment belonging to a particular cluster. We determine cluster membership for the segments based on the frequency of the A allele at each locus and the probability that a transition between clusters occurs between two loci as a function of the distance between the two loci.

Similar to fastPHASE [56], Bayes IM uses the maximum likelihood estimates from a hidden Markov Model in order to estimate the emission and transition probabilities. Bayes IM uses a fixed number of haplotype cluster, N, which are unobserved hidden states in a hidden Markov model. For each individual i and locus k we can observe their genotype, x_{ik}, which is emitted from a pair of unordered haplotype clusters, $\pi_{ik} = \{S_l, S_m\}$.

We can define the initial state probabilities as the probability that at locus 1 individual i begins in the unordered haplotype cluster pair $\pi_{i1} = \{S_l, S_m\}$. This is denoted by:

$$P\left(\pi_{i1} = \{S_l, S_m\}\right) = \begin{cases} \left(\alpha_{S_l 1}\right)^2 & S_l = S_m \\ 2\alpha_{S_l 1}\alpha_{S_m 1} & S_l \neq S_m \end{cases},$$

where $\alpha_{S_l 1}$ is the probability that, at locus 1, we are in haplotype cluster S_l. The probability that individual i transitions from cluster S_l at locus $k-1$ to cluster S_m at locus k on a chromosome is defined as follows:

$$a\left(k\right)_{S_l S_m} = P_k\left(S_l \rightarrow S_m\right) = P_k\left(S_m \in \pi_{ik} | S_l \in \pi_{i(k-1)}, \lambda\right)$$

$$= \begin{cases} e^{-\frac{d_k}{\lambda}} & No\, transition\, occurs \\ \left(1 - e^{-\frac{d_k}{\lambda}}\right)(\alpha_{S_m k}) & Transition\, occurs \end{cases},$$

where d_k is the physical distance between locus $k-1$ and k and λ is a parameter which needs to be estimated. Defining the transition probability in this way establishes two cases. The case where no transition occurs, or we remain in the same state, and the case where a transition occurs. We should also note that it is possible for us to transition back to the same state we started in. The emission probability, $e_{\pi_{ik}}(x_{ik})$, is the probability that individual i at locus k has observed genotype x_{ik} given we are in cluster pair $\pi_{ik} = \{S_l, S_m\}$. The emission probability is written:

$$e_{\pi_{ik}}(x_{ik}) = P_k(x_{ik}|\pi_{ik} = \{S_l, S_m\}, \theta)$$

$$= \begin{cases} (1 - \theta_{S_l k})(1 - \theta_{S_m k}) & x_{ik} = 0 \\ \theta_{S_l k}(1 - \theta_{S_m k}) + \theta_{S_m k}(1 - \theta_{S_l k}) & x_{ik} = 1 \\ \theta_{S_l k}\theta_{S_m k} & x_{ik} = 2 \end{cases},$$

where $\theta_{S_l k}$ is the allele frequency of the A allele in cluster S_l at locus k [56].

2.5.5.2 Develop the Sampler

Bayes IM differs from previous haplotype models. Previous models use the HMM to phase the genotypes and then treats the estimated haplotypes as known. Bayes IM instead uses the HMM as a part of the model. Bayes IM uses a Markov chain Monte Carlo (MCMC) sampler to first sample the haplotype clusters, then sample for a QTL and cluster effects at each locus, and finally sample the fixed effects, random effects, and variances. Thus haplotype

clusters are sampled on an individual basis and are re-sampled with each iteration rather than being fixed as they are in previous haplotype models. We can sample the haplotype clusters for individual i with genotype sequence X_i using the probability of observing haplotype cluster sequence π_i defined by $P(\pi_i \mid X_i) \propto P(X_i, \pi_i)$, which was defined above in section 2.5.3.

Now recall, we defined the model equation as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e},$$

or equivalently:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZM_H}\boldsymbol{b} + \mathbf{e}.$$

We define $\mathbf{M_H}$ as an $n \times K_H$ matrix of haplotype covariates, where n is the number of individuals, $K_H = N \times k$ is the total number of haplotype effects which need estimated, N is the number of haplotypes in each block, which is fixed, and k is the number of haplotype loci. For individual i, haplotype locus h, and cluster c the haplotype covariate is:

$$m_{ihc} = \begin{cases} 0 & \text{if } S_l \neq c \text{ and } S_m \neq c \\ 1 & \text{if } S_l = c \text{ and } S_m \neq c \text{ or } S_l \neq c \text{ and } S_m = c \\ 2 & \text{if } S_l = S_m = c \end{cases},$$

where $\pi_{ik} = \{S_l, S_m\}$. The haplotype cluster effect vector $\mathbf{b} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_H\}$ is a K_H vector of random regression coefficients of the K_H haplotypes, where $\mathbf{b}_h$ represents a vector of size N random regression coefficients for the N haplotype clusters in haplotype block h. Similar to Bayes C [37], at haplotype locus h, the haplotype cluster vector $\mathbf{b}_h$ will be $\mathbf{0}$ with probability of π and normally

distributed with a mean of $\mathbf{0}$ and variance of $\mathbf{I}\sigma_b^2$ with probability $1-\pi$, which we denote as:

$$\mathbf{b}_h|\sigma_b^2 \sim \begin{cases} \mathbf{0} & \text{with probability } \pi \\ N(\mathbf{0},\mathbf{I}\sigma_b^2) & \text{with probability } 1-\pi \end{cases}.$$

The fixed effects, random effects, and variances are sampled just like they were in the Bayes C model. The fixed effects, $\boldsymbol{\beta}$, are assigned an improper flat prior, which is denoted as:

$$\boldsymbol{\beta} \sim U(-\infty, \infty).$$

The random effects, $\mathbf{e}$, have a multivariate normal prior distribution with a mean of $\mathbf{0}$ and a covariance matrix of $\mathbf{R}\sigma_e^2$, where $\mathbf{R}$ is a diagonal matrix. We will denote this as:

$$\mathbf{e}|\mathbf{R}\sigma_e^2 \sim N(\mathbf{0}, \mathbf{R}\sigma_e^2).$$

For the variances, σ_b^2 and σ_e^2 we use a scaled inverse chi-squared prior distribution with known scale parameter, s_b^2 and s_e^2, and known degrees of freedom, ν_b and ν_e, which is denoted as:

$$\sigma_b^2|\nu_b, s_b^2 \sim Inv-\chi^2(\nu_b, s_b^2)$$

and

$$\sigma_e^2|\nu_e, s_e^2 \sim Inv-\chi^2(\nu_e, s_e^2)$$

respectively [37].

CHAPTER 3

ACCOUNTING FOR A COMPOSITE BREED POPULATION

3.1 Introduction

Current methods of genomic prediction used in the livestock industry focus on improving predictability within purebred populations by utilizing the assumption that individuals are being sampled from an homogeneous population. The homogeneity assumption is often violated in a commercial setting since commercial livestock populations are often composites of multiple breeds [50]. When the current models for genomic prediction are used it can result in predictors that perform poorly when used to predict the relative genetic merit of individuals whose breed composition differs from the training population. An explanation for this poor performance is due to changes in linkage disequilibrium between the markers and the quantitative trait loci depending on which breed the chromosomal segment originated in.

There have been few studies which analyzed the effect on prediction accuracy when a composite breed population was used. Bolormaa et al. [2013] showed that with composite breed populations prediction accuracy is improved when you include composite individuals and pure bred individuals together in the training set versus only including breed specific individuals. Studies using pure breed populations can also give us some insight. Hayes et al. [2009] and

Kachman et al. [2013] used pure breed populations to demonstrate that, in most cases, prediction accuracy is improved when the breed with which you wish to make evaluations on is included in the training set. Muijibi et al. [2011] showed that prediction accuracy was improved when only one breed was used for both training and evaluation rather than using all the breeds together for training and evaluation.

Similar to other GWAS models Bayes IM, proposed by Kachman [2015], fails to account for the breed composition of the individual. The proposed solution is to adapt Bayes IM to include breed specific haplotype clusters rather than using a common set of haplotype clusters. This adapted model will be called Bayes IM Comp.

3.2 Description of Bayes IM Comp

Now, we want to consider a population of individuals made up of B breeds. To model differences in haplotype clusters between the B breeds, the N haplotype clusters are partitioned into B sets of breed specific haplotype clusters. For breed b, it is assumed there are N_b haplotype clusters contained in its haplotype cluster group. The total number of haplotype clusters is $N = \sum_b N_b$.

Let $C_i = \{C_{i1}, C_{i2}, ..., C_{iB}\}$ be the breed composition of individual i, where C_{ib} is the proportion of individual i which is from breed b. The probability that, at locus k, we are in haplotype cluster S_{l_b} given we are in haplotype cluster group b is:

$$\alpha_{S_{l_b}k} = \frac{1}{N_b}.$$

That is within a particular breed each haplotype is assumed to be equally likely. The probability that, for individual i, at locus k the haplotype cluster came from group b is assumed to be equal to C_{ib}. Thus, the probability that,

at locus k, individual i is in haplotype cluster S_{l_b} is

$$C_{ib} \cdot \alpha_{S_{l_b}k} = C_{ib} \cdot \frac{1}{N_b}.$$

Initial state probabilities and transition probabilities are as follows. The probability, at locus 1, of individual i beginning in unordered haplotype cluster pair $\pi_{i1} = \{S_{l_b}, S_{m_c}\}$ is defined to be:

$$a(1)_{\pi_{i1}} = P_1(\pi_{i1} = \{S_{l_b}, S_{m_c}\}) = \begin{cases} \left(C_{ib}\alpha_{S_{l_b}1}\right)^2 & S_{l_b} = S_{m_c} \\ 2\left(C_{ib}\alpha_{S_{l_b}1}\right)\left(C_{ic}\alpha_{S_{m_c}1}\right) & S_{l_b} \neq S_{m_c} \end{cases}.$$

The probability that individual i transitions on a chromosome from cluster S_{l_b} in breed b at locus $k-1$ to cluster S_{m_c} in breed c at locus k is defined by

$$a(k)_{S_{l_b}S_{m_c}} = \begin{cases} e^{-d_k r} & No\,transition\,occurs \\ \left(1 - e^{-d_k r}\right)\left(C_{ic} \cdot \frac{1}{N_c}\right) & Transition\,occurs \end{cases}$$

where, d_k is the physical distance between markers $k-1$ and k and $r = \frac{1}{\lambda}$ is a parameter which needs to be estimated. The remaining components of this model match the Bayes IM model described previously.

3.3 Accounting for Within and Between Breed Transitions in Bayes IM Comp

Bayes IM Comp uses a single scale parameter λ to define the transition probability between two loci. Thus we are equally likely to jump to a haplotype cluster in another breed as we are to jump to a haplotype cluster within the same breed. As transitions between clusters within the same breed versus

transitions between clusters across different breed may represent recombination events on different time scales, having λ vary between and across breeds may be reasonable. We achieve this by defining two rate parameters r_B and r_W, where $r_B + r_W = r = \frac{1}{\lambda}$. The rate parameter, r_B, defines the frequency of transitions between haplotype clusters across breeds and the rate parameter, r_W, defines the frequency of transitions within haplotype clusters of the same breed.

When accounting for between breed and within breed haplotype cluster transitions, we can define three separate cases, whereas before we only defined two. The first case is where no transition occurs and we remain in the same state. The second case is where a transition occurs within haplotype clusters of the same breed, but we can also transition back to the same cluster we started in. The last case is where a transition occurs between haplotype clusters of different breeds; however, we can still transition across breeds and return to a cluster within the starting breed or the starting cluster itself. The probability that individual i transitions from haplotype cluster S_{l_b} within breed b to haplotype cluster S_{m_c} within breed c at locus k is denoted by:

$$a\left(k\right)_{S_{l_b}S_{m_c}} = \begin{cases} e^{-d_k r_W} e^{-d_k r_B} & No\, transition\, occurs \\ e^{-d_k r_B}\left(1 - e^{-d_k r_W}\right)\left(\frac{1}{N_c}\right) & Transition\, within\, breed\, occurs \\ \left(1 - e^{-d_k r_B}\right)\left(C_{ic} \cdot \frac{1}{N_c}\right) & Transition\, across\, breeds\, occurs \end{cases} .$$

3.4 Utilizing Breed composition information from the parents

An individuals genotype is made up of two haplotypes. One haplotype is inherited from the mother while the other haplotype is inherited from the

father. We can utilize the breed composition information from the parents rather than the individual in order to better control the possible haplotype cluster pairs for an individual. We call this model Bayes IM Parental Comp or Bayes IM PC for short. This model would be appropriate for an individual in which their mother comes from one breed and their father is from another breed.

As before, assume we have a population of individuals made up of B breeds. We assume that each breed, b, is made up of N_b haplotype clusters. Similar to Bayes IM Comp above, the overall number of haplotype clusters is equal to $\sum_b N_b = N$. Let $M_i = \{M_{i1}, ..., M_{iB}\}$ be the maternal breed composition for individual i and $P_i = \{P_{i1}, ..., P_{iB}\}$ be the paternal breed composition for individual i. M_{ib} and P_{ib} are the proportion of individual i's mother and father, respectively, which comes from breed b.

As mentioned above, an individual's genotype is made up of two haplotypes. While we can observe the genotype for individual i at locus k, we cannot observe the actual haplotype clusters which are generating this genotype. Unlike Bayes IM and Bayes IM Comp above, π_{ik} denotes the ordered pair of clusters with which genotype x_{ik} for individual i at locus k originates. Order matters here since we need to distinguish which haplotype cluster is inherited from the mother and which is inherited from the father. We let $\pi_{ik} = \{z(m)_{ik}, z(p)_{ik}\}$, where $z(m)_{ik}$ represents the haplotype cluster coming from the mother and $z(p)_{ik}$ represents the haplotype cluster coming from the father. We can now define the initial state probabilities and transition probabilities which make up the hidden Markov model.

At locus k, the probability that individual i's maternal haplotype cluster

is S_{l_b} is:

$$P\left(z\left(m\right)_{ik} = S_{l_b}\right) = M_{ib} \cdot \alpha_{S_{l_b}k} = M_{ib} \cdot \left(\frac{1}{N_b}\right),$$

where $\alpha_{S_{l_b}k}$ is the probability that at locus k haplotype cluster S_{l_b}occurs given we are in a haplotype cluster from breed b. We assume that all haplotype clusters within a breed are equally likely, therefore, $\alpha_{S_{l_b}k} = \frac{1}{N_b}$. Similarly, at locus k, the probability that individual i's paternal haplotype cluster is S_{m_c} is:

$$P\left(z\left(p\right)_{ik} = S_{m_c}\right) = P_{ic} \cdot \alpha_{S_{m_c}k} = P_{ic} \cdot \left(\frac{1}{N_c}\right).$$

The initial state probabilities are defined at follows. At locus 1, the probability that individual i's maternal haplotype cluster is S_{l_b} and individual i's paternal haplotype cluster is S_{m_c} is defined by:

$$P\left(\pi_{i1} = \left\{z\left(m\right)_{i1}, z\left(p\right)_{i1}\right\}\right) = P\left(\pi_{i1} = \left\{S_{l_b}, S_{m_c}\right\}\right) = \left(M_{ib} \cdot \frac{1}{N_b}\right) \cdot \left(P_{ic} \cdot \frac{1}{N_c}\right).$$

The transition probabilities can be broken up into the maternal haplotype transition and the paternal haplotype transition. At locus k, the probability that the maternal haplotype cluster is S_{m_c}, given that at locus $k-1$ the maternal haplotype cluster was S_{l_b} is:

$$a\left(m,k\right)_{S_{l_b}S_{m_c}} = P_{m,k}\left(S_{l_b} \to S_{m_c}\right) = P\left(z\left(m\right)_{ik} = S_{m_c} \mid z\left(m\right)_{i(k-1)} = S_{l_b}\right)$$

$$= \begin{cases} e^{-d_k r_W} e^{-d_k r_B} & No\,transition\,occurs \\ e^{-d_k r_B}\left(1 - e^{-d_k r_W}\right)\left(\frac{1}{N_c}\right) & Transition\,within\,breed\,occurs \\ \left(1 - e^{-d_k r_B}\right)\left(M_{ic} \cdot \frac{1}{N_c}\right) & Transition\,across\,breeds\,occurs \end{cases} \cdot$$

Similarly, at locus k, the probability that the paternal haplotype cluster is

$S_{m'_{c'}}$, given that at locus $k-1$ the paternal haplotype cluster was $S_{l'_{b'}}$ is:

$$a\left(p,k\right)_{S_{l'_{b'}}S_{m'_{c'}}} = P_{p,k}\left(S_{l'_{b'}} \rightarrow S_{m'_{c'}}\right) = P\left(z\left(p\right)_{ik} = S_{m'_{c'}} \mid z\left(p\right)_{i(k-1)} = S_{l'_{b'}}\right)$$

$$= \begin{cases} e^{-d_k r_W} e^{-d_k r_B} & No\, transition\, occurs \\ e^{-d_k r_B}\left(1 - e^{-d_k r_W}\right)\left(\frac{1}{N_{c'}}\right) & Transition\, within\, breed\, occurs \\ \left(1 - e^{-d_k r_B}\right)\left(P_{ic'} \cdot \frac{1}{N_{c'}}\right) & Transition\, across\, breeds\, occurs \end{cases} .$$

Then the probability that at locus $k-1$ we were in haplotype clusters $\pi_{i(k-1)} = \left\{S_{l_b}, S_{l'_{b'}}\right\}$ and transition to haplotype clusters $\pi_{ik} = \left\{S_{m_c}, S_{m'_{c'}}\right\}$ at locus k is:

$$P\left(\pi_{i(k-1)} \rightarrow \pi_{ik}\right) = P\left(\pi_{ik} = \left\{S_{m_c}, S_{m'_{c'}}\right\} \mid \pi_{i(k-1)} = \left\{S_{l_b}, S_{l'_{b'}}\right\}\right)$$

$$P\left(z\left(m\right)_{ik} = S_{l_b}, z\left(p\right)_{ik} = S_{l'_{b'}} \mid z\left(m\right)_{i(k-1)} = S_{m_c}, z\left(p\right)_{i(k-1)} = S_{m'_{c'}}\right)$$

$$P\left(z\left(m\right)_{ik} = S_{m_c} \mid z\left(m\right)_{i(k-1)} = S_{l_b}\right) \cdot P\left(z\left(p\right)_{ik} = S_{m'_{c'}} \mid z\left(p\right)_{i(k-1)} = S_{l'_{b'}}\right).$$

CHAPTER 4

EVALUATION OF BAYES IM COMP

4.1 Introduction

In livestock, prediction of an individual's genetic merit using genomic information is becoming increasingly common. The models used to make these predictions typically assume that we are sampling from a homogeneous population. However, in both commercial and experimental populations the sire and dam of an individual may be a mixture of several populations. Bayes IM Comp is based on breed specific haplotype clusters which utilizes the known breed composition from the individual and was developed to allow for differences in linkage disequilibrium across multiple breeds. Bayes IM Comp will be compared to the Bayes B and C models and to the Bayes IM model using two separate populations below. We will be evaluating the models based on their ability to detect a QTL and on their prediction ability. The first population is a composite beef cattle population strongly populated with Simmental genetics. The second population is a simulated composite cattle population. Since the true location and size of the QTLs are known for the simulated data set, we will be able to make a more accurate assessment of the ability of each model to detect the true QTLs.

4.2 Evaluation of Bayes IM Comp Using the Simmental Data Set

4.2.1 Data Description

Genotypes on 6,552 Simmental and Simmental composite cattle from the American Simmental Association were used. The genotypes consisted of a total of 99,827 mapped autosomal SNPs from two different genotyping platforms, in which 27,562 were common between the two platforms. The average breed composition of the genotyped individual was 63% Simmental, 30% Angus, and 2% Hereford. With the remaining 29 breeds together accounting for the other 5%. Thus, we considered the percentage of each individual which came from a Simmental, Angus, Hereford, and combined breed background. Expected progeny differences (EPDs) for the following traits were evaluated. Traits evaluated include five weight traits including birth weight (BWT), weaning weight due to milk (MILK), weaning weight due to both milk and growth (MWWT), direct weaning weight (WWT) or weaning weight due to growth, and yearling weight (YWT). In addition, five carcass traits were evaluated including carcass weight (CWT), back fat (BFAT), marbling score (MARB), ribeye muscle area (REA), and yield grade (YG). Calving ease (CE), docility (DOC), and maternal calving ease (MCE) were three threshold traits evaluated.

4.2.2 Models Compared

For each trait six models were considered: Bayes B, Bayes C, Bayes IM 16 (100% common model, this is the Bayes IM model with a total of 16 haplotype clusters shared by all breeds), Bayes IM 8 (100% common model, this is the Bayes IM model with a total of 8 haplotype clusters shared by all breeds),

Bayes IM Comp 50 (50% common, this is the Bayes IM Comp model with 4 haplotype clusters assigned to the Simmental breed, 2 assigned to Angus, 1 to Hereford, 1 to the combined breed group, and 8 assigned to be common among all breeds for a total of 16 haplotype clusters), and Bayes IM Comp (0% common, this is the Bayes IM Comp model with 8 haplotype clusters assigned to the Simmental breed, 4 assigned to Angus, 2 to Hereford, and 2 to the combined breed group, and zero haplotype clusters assigned to be common among all breeds for a total of 16 haplotype clusters).

The EPDs for each trait were deregressed to account for variable accuracies [16]. The deregression was carried out assuming 40% of the genetic variability was due to polygenic effects. In addition to the SNP effects, the model included an overall mean as a fixed effect. The residual for individual i was assumed to be $N\left(0, \frac{\sigma_e^2}{w_i}\right)$, where w_i is a weight to account for differences in the residual variance and σ_e^2 was scaled to be 1.5 times the genetic variance, which corresponds to a heritability of 0.4.

4.2.3 Training and Evaluation Sets

All individuals who had a SNP genotype and breed composition were included in the analysis. Table 4.2.1 shows the number of individuals remaining after removing individuals missing EPDs for a particular trait. First, a random two-thirds of the 3,752 individuals from the direct weaning weight trait were used to train the parameters of the HMM model. This ensured that we had a good mix of all possible breed compositions in order to estimate the breed specific haplotype clusters and estimate the transition parameters r_B and r_W for the Bayes IM Comp models and r for the Bayes IM model. After the parameters for the HMM were estimated, three folds were created by partitioning

Table 4.2.1: Number of individuals used in the training and evaluation sets per trait

Trait	High Fold	Medium Fold	Low Fold	Total
BWT	1681	1271	913	3865
MILK	1488	1165	860	3513
MWWT	1556	1225	894	3675
WWT	1590	1258	904	3752
YWT	1600	1260	903	3763
CWT	1613	1272	900	3785
BFAT	953	974	589	2516
MARB	913	958	583	2454
REA	883	938	572	2393
YG	989	980	594	2563
CE	1649	1254	901	3804
DOC	1149	677	365	2191
MCE	1536	1218	889	3643

the data into a fold of individuals with high Simmental breed composition, a fold of individuals with medium Simmental breed composition, and a fold of individuals with low Simmental breed composition. On average, the high Simmental fold had a breed composition of 85% Simmental, the medium fold was 66% Simmental, and the low fold was 40% Simmental.

4.3 Results and Discussion for the Simmental Data Set

Since many of the trends are similar between all the traits, we will only examine two carcass traits and two weight traits in detail when assessing QTL identification and haplotype effect estimates. The two carcass traits are ribeye area and yield grade and the two weight traits are direct weaning weight and yearling weight. For the posterior distribution comparisons we will only be examining yearling weight in detail. The results for the other three traits can be found in Appendix A. All 13 traits were used to evaluate prediction accu-

racy. However, the primary question we want to answer is how will each model work for individuals with the smaller percentage of the Simmental breed. Thus we will only show the prediction accuracy results for the low Simmental fold. Additional results can be found in Appendix A.

4.3.1 Posterior Distribution Comparison

The first step is to examine the prior and posterior values for the parameters. Priors for each trait were estimated assuming a heritability of 0.4, consistent with the scaling of σ_e^2 during the deregression of the EPDs. Thus, the residual variance is equal to 60% of the sample variance of the EPDs and the genetic variance is equal to 40% of the sample variance of the EPDs. The haplotype effect variance was estimated to be:

$$\sigma_b^2 = \frac{\sigma_g^2}{\left((1-\pi)\cdot n_{QTL}\cdot\left(1-\frac{1}{N_{cluster}}\right)\cdot 2\right)},$$

where σ_g^2 is the estimated genetic variance, π ($= 0.975$) is the proportion of markers which have no effect, n_{QTL} ($\approx 10,000$) is the number of putative QTL in the model, and $N_{cluster}$ ($= 16$) is the number of haplotype clusters in the model. The Bayes IM 8 model used the same haplotype effect variance prior as the Bayes IM 16 model since the difference between using $N_{cluster} = 16$ and $N_{cluster} = 8$ was small. No fine tuning of the prior estimates were preformed on this data set.

Table 4.3.1 displays the prior estimates, posterior means, and standard deviations for the YWT trait. The posterior means for REA, YG and WWT can be found in Appendix A Tables A.1.1, A.1.2 and A.1.3, respectively.

For all four traits examined, the residual variance prior is higher than

Table 4.3.1: YWT: Prior and Posterior Means (SE) for Variance Components

Model	Genetic Variance	Residual Variance	Heritability	Haplotype Effect Variance
Prior	663	995	0.4	1.3
Bayes B	667.4 (26.64)	689.1 (23.57)	0.492 (0.016)	N/A
Bayes C	736.9 (34.73)	639.2 (25.85)	0.535 (0.019)	N/A
Bayes IM 16	971.6 (44.62)	445.1 (30.28)	0.686 (0.023)	2.10 (0.199)
Bayes IM 8	951.0 (48.45)	459.7 (32.97)	0.674 (0.026)	2.09 (0.184)
Bayes IM Comp 50	1060.7 (49.45)	375.4 (33.18)	0.738 (0.025)	2.25 (0.212)
Bayes IM Comp	1161.9 (67.35)	300.8 (47.25)	0.794 (0.036)	2.52 (0.244)

the posterior mean for all six models. The Bayes B and Bayes C models are estimating the residual variance to be slightly higher than any of the Bayes IM models. In addition, the Bayes IM Comp models are estimating the residual variance to be lower than that estimated by the original Bayes IM models.

The general observed pattern for the genetic variance is that Bayes B and C have the lowest posterior means for the genetic variance and the Bayes IM Comp models have the highest posterior means. This is consistent with the trend observed within the residual variance. As the estimated genetic variance increases, the estimated residual variance decreases since the overall variance is being shared between the genetic and residual variances.

Overall, all the posterior mean heritability estimates are greater than the 0.4 value used to estimate the prior values. The higher residual variance values and lower genetic variance values seen in the Bayes B and C models leads to these two models producing lower heritability estimates versus the Bayes IM models. Similarly, the lower residual variance values and higher genetic variance values observed from the Bayes IM Comp models produces the observed higher heritability.

The posterior means for Bayes B are closest to the prior values. This is because Bayes B estimates a locus specific variance which makes the model more sensitive to the prior information. The estimates for the four haplotype

based Bayes IM models are much further away from the prior values than either Bayes B or Bayes C. One explanation for this is that the Bayes IM models are not influenced as strongly by the prior information as Bayes B or Bayes C. A second possible explanation is that the haplotypes in the Bayes IM model are doing a much better job of capturing the true genetic variance than the SNPs are in Bayes B and Bayes C.

Since Bayes B and C are not haplotype models, they have no haplotype effect variance; however they do have a SNP effect variance which was not reported. Comparing just the four Bayes IM models, we observe the same pattern seen with the genetic variance for the haplotype effect variance. The original Bayes IM models have a lower haplotype effect variance than the Bayes IM Comp models. Since the haplotype effect variance is a function of the genetic variance this pattern make sense.

In addition to the posterior means, the posterior distributions of the parameters were examined and can be found in Appendix A. For YWT, the density plots for Bayes B and Bayes C can be found in Figure A.1.7 and the density plots for the Bayes IM models can be found in A.1.8. Ideally, we want to see density plots which are symmetric and bell shaped. We do not want to see skewed distributions as this is an indication that the initial burn-in was too small and we should discard a larger number of MCMC iterations in order to give the model time to settle in. Based on the density plots, we conclude that the model had a large enough burn-in to settle in since the plots all appear to be mostly symmetric and bell shaped. None of the posterior distributions are centered at the prior values as indicated by the posterior mean values, but this did not concern us as our priors are not skewing the posterior distribution.

A second method to ensure that the initial burn-in is large enough is to

examine the trace plots. The trace plots were examined and overall appeared random. No obvious patterns were observed, which indicates that our burn-in is sufficient enough for a models to stabilize. Additionally, this is an indication that the overall number of MCMC samples is large enough to get a good estimate of the posterior for each parameter.

4.3.2 QTL Identification and Haplotype effect estimates

Table 4.3.2: Documented QTLs Associated with More than One Trait or Segregating in More than One Breed

BTA_MB	Start	Stop	Associated Traits	Breeds
2_6	6,047,202	6,831,955	REA,WWT,YG	LIM
5_48	48,080,258	48,993,294	REA	ANG,BRG
5_106	106,156,727	106,977,557	WWT,YWT	HH
6_38	38,042,010	38,939,012	REA,WWT,YWT	GVH,HH,LIM,RAN,SIM
6_42	42,023,748	42,906,960	YG	RDP
7_93	93,007,434	93,886,136	REA,WWT,YWT	ANG,HH,SIM
14_24	24,057,353	24,787,245	WWT,YWT	GVH,SIM
15_38	38,003,133	38,957,876	REA	HH
20_4	4,043,932	4,989,460	WWT,YG,YWT	ANG,HH,RAN,SIM

a. Table based on Saatchi et al. (2014) [52]
b. ANG (Angus), BRG (Brangus), GVH (Gelbvieh), HH (Hereford), LIM (Limousin), RAN (Red Angus), RDP (Maine-Anjou), SIM (Simmental)

The second step is to compare the 6 models ability to identify a QTL. QTL identification will be based on the genetic variance for each SNP in the Bayes B and C models and each putative QTL in the Bayes IM versions. This will allow us to see where the genetic variance peaks which indicates the presence of a QTL. Saatchi et al. [52] identified several large-effect QTL which are associated with several traits or are segregating within several breeds. These QTLs were 1 MB windows which explained more the 1% of the additive genetic variance. Table 4.3.2 summarizes the nine QTLs associated with REA, YG, WWT, and YWT. Similar to Saatchi et al., we broke the genome up into 1

MB windows and, for each trait, calculated the genetic variance within that window and ranked all the windows from largest to smallest. We looked at the windows which were 1 MB below, 1 MB above, and at the identified QTL and if the rank of the window was less than 100, we said that the model identified the QTL. If the rank was between 101 and 200 we said the model nearly identified the QTL. If the rank was greater than 200 then we said that the model did not identify the QTL. It should be noted that there is more than one QTL on BTA 6, BTA 14, and BTA 20; however, we chose to report only one QTL in Table 4.3.2.

4.3.2.1 REA

We will first examine the QTLs for the carcass trait of REA, where REA is a measure of the size of the ribeye muscle at the 12th rib. A summary of the identified QTLs for each of the six models can be found in Table 4.3.3. Of the five QTLs associated with REA, the four Bayes IM models identified three. Bayes C only identified two of the QTL and Bayes B identified two and nearly identified one. The QTL on BTA 2 and BTA 15 were not identified by any of the models. It should be noted that the QTL on BTA 2 is only segregating within the Limousin breed which is part of the combined breed group and accounts for approximately 5% of the genetic information. Additionally, BTA 15 is only segregating within the Hereford breed which accounts for only 2% of the genetic information. In order to better detect these QTL, we need more individuals with genetics coming from the Limousin and Hereford breeds. The QTL on BTA 5 at 48 MB is segregating within the Angus and Brangus breeds. Angus accounts for 30% of the genetic information in the Simmental data set. The Bayes IM models which utilize haplotypes appear to better detect this

Table 4.3.3: REA: QTLs Identified in the top 100 1 MB Windows

BTA_MB	Bayes B	Bayes C	Bayes IM 16	Bayes IM 8	Bayes IM Comp 50	Bayes IM Comp
2_6	No	No	No	No	No	No
5_48	No** (167)	No	Yes (8)*	Yes (85)*	Yes* (37)	Yes* (30)
6_38	Yes (17)	Yes (11)	Yes (89)	Yes (4)	Yes (10)	Yes (40)
7_93	Yes (65)	Yes (26)	Yes (2)	Yes (2)	Yes (2)	Yes (4)
15_38	No	No	No	No	No	No

a. * MB below, **MB above

QTL.

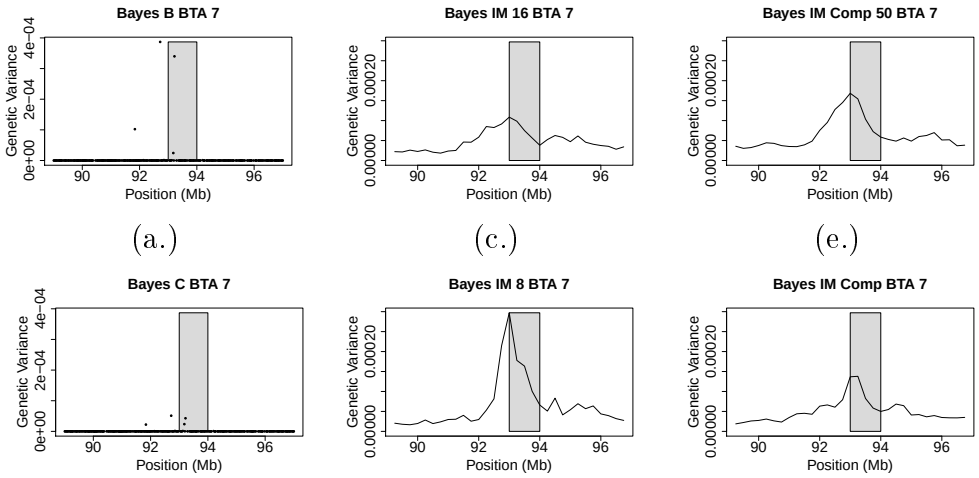
Rather than look at all five locations in detail here, we will only look at one location. The details for the other four locations can be found in Appendix A. The location we will examine in detail is BTA 7 at 93 MB. This location is segregating in the Simmental, Angus, and Hereford breeds which are the three main breed groups for this data set. The Bayes IM models all rank this QTL window in the top 5, Bayes C ranks it 26th and Bayes B ranks it 65th.

The individual SNP genetic variances for Bayes B and Bayes C are plotted in Figure 4.3.1 (a.) and (b.), respectively. The individual putative QTL genetic variances for the Bayes IM models are plotted in Figure 4.3.1 (c.) through (f.). Even though Bayes B had to lowest rank of all the models examined, it is showing several SNPs with a genetic variance that are much larger than that seen in Bayes C. These large SNPs observed within Bayes B model are due to Bayes B using a t-distribution with heavy tails which causes SNPs in the tail of the distribution to have a larger effect. Bayes C has several SNPs with a small elevation in genetic variance, but the QTL is not as definitive as the QTL in Bayes B.

Of all the Bayes IM models, Bayes IM 8 is showing the largest peak, followed by Bayes IM Comp 50. Bayes IM 16 and Bayes IM Comp have peaks of similar magnitude but Bayes IM Comp has a much narrower QTL peak. Since Bayes B and Bayes C are on a SNP level and the Bayes IM models are

on a haplotype block level there is no true way for us to compare these models other than by comparing their ranks.

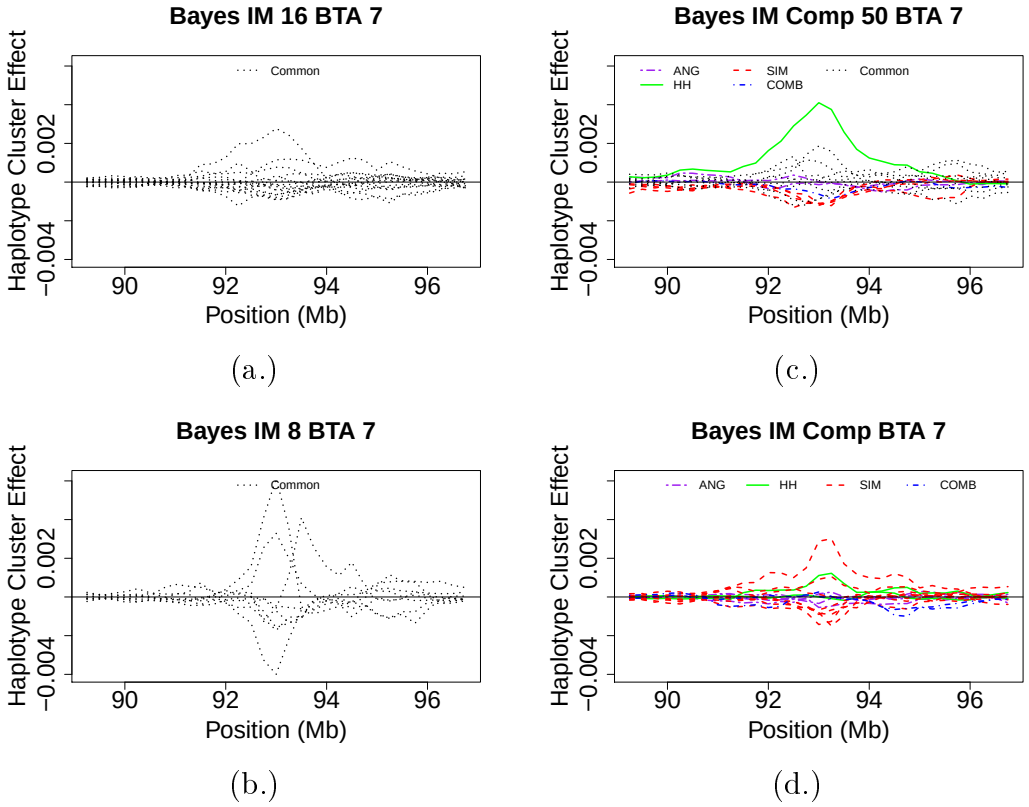
Figure 4.3.1: REA: Genetic Variance for BTA 7 between 89 and 97 MB



a. The shaded region indicates the area with which we expect to observe peaks in the genetic variance.

We will now examine the haplotype cluster effects for the four Bayes IM models, which can be seen in Figure 4.3.2. Bayes IM 8, which had the largest genetic variance out of these four models, is showing three haplotype clusters with a large positive effect and one haplotype cluster with a large negative effect. Bayes IM 16, on the other hand, is only showing one haplotype cluster with a large positive effect and zero haplotype clusters with a large negative effect. What is happening between these two models is that effects for the clusters in Bayes IM 8 are being spread among multiple haplotype clusters within Bayes IM 16, which results in the muted effects we are observing within Bayes IM 16. For example, haplotype clusters 4, 12, and 15 within Bayes IM 16 have small positive effects. Additionally, haplotype cluster 16 within Bayes IM 16 is the cluster with the largest positive effect for Bayes IM 16. We calculated the probability that an individual belongs to a particular cluster within Bayes

Figure 4.3.2: REA: Haplotype Effect Estimates for BTA 7 Between 89 and 97 MB



IM 16 given they were in haplotype cluster 1 within Bayes IM 8, which is one of the large positive effect clusters from Bayes IM 8. This probability revealed that 33% of individuals are placed in haplotype cluster 16 within Bayes IM 16. Additionally, 8% were placed in haplotype cluster 4, 6% in haplotype cluster 12, and 14% in haplotype cluster 15 within Bayes IM 16. This is causing the large effect from Bayes IM 8 to be muted within Bayes IM 16.

Bayes IM Comp is showing one Simmental haplotype cluster with a large positive effect, a Hereford and Simmental haplotype cluster with a medium positive effect, and several Simmental haplotype clusters with a medium negative effect. None of the Angus haplotype clusters are having an effect. Bayes IM Comp 50 has a Hereford haplotype cluster with a large positive effect,

several common haplotype clusters with a medium positive effect, and several Simmental and common haplotype clusters with a small negative effect. In order to examine the large Hereford cluster from Bayes IM Comp 50, we calculated the probability that an individual belongs to a particular cluster within Bayes IM Comp given there were in the Hereford cluster within Bayes IM Comp 50. We discovered that individuals who were in the Hereford cluster in Bayes IM Comp 50 had a probability of 50% of being in Simmental cluster 7, a 10% probability of being in Simmental cluster 8, and a 6% probability of being in either Simmental cluster 13 or cluster 14. This explains why the large effect of the Hereford cluster in Bayes IM Comp 50 is seems to disappear in Bayes IM Comp. The effects from the Hereford cluster in Bayes IM Comp 50 is being spread among several clusters within Bayes IM Comp. Additionally, this is an indication that our models are not properly identifying the breed specific haplotype clusters.

4.3.2.2 YG

Table 4.3.4: YG: QTLs Identified in the top 100 1 MB. Windows

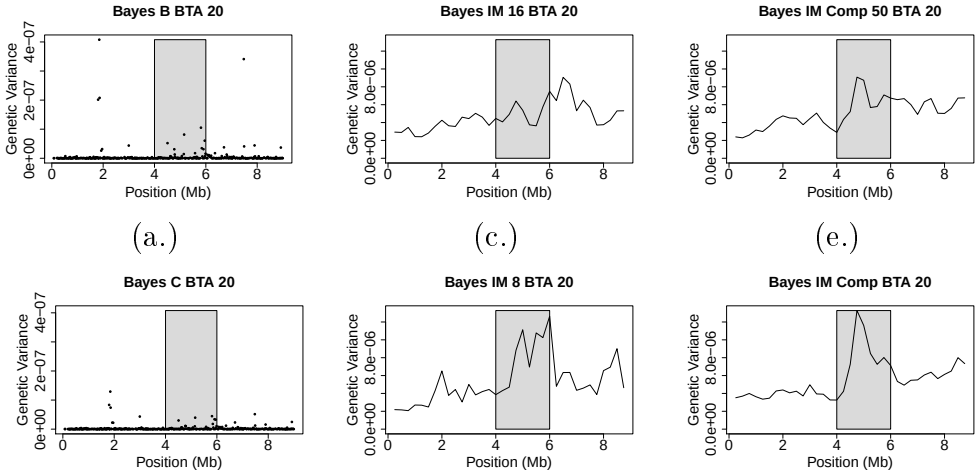
BTA_MB	Bayes B	Bayes C	Bayes IM 16	Bayes IM 8	Bayes IM Comp 50	Bayes IM Comp
2_6	Yes (93)	Yes (51)	Yes (57)	No	Yes (19)	Yes (64)
6_42	No	No	No	No	No	No
20_4	No** (139)	No	No	Yes** (53)	No** (149)	Yes** (12)

a. * MB below, **MB above

The second carcass trait we will examine for QTLs is YG, which refers to the amount of usable meat obtained from a carcass. There were only three identified QTLs for this trait and Table 4.3.4 summarizes which of the six models was able to currently identify each of the QTLs. The QTL on BTA 2 at 6 MB, which is associated with the Limousin breed, was identified by every model except for the Bayes IM 8 model. The QTL on BTA 6 at 42 MB was

not identified by any of the models. The QTL on BTA 6 is segregating only within the Maine-Anjou breed which is not one of the major breed groups for this data set. Finally, the QTL on BTA 20 at 4 MB was identified by Bayes IM 8 and Bayes IM Comp at 5 MB. Bayes IM Comp 50 and Bayes B nearly identified this QTL at 5 MB and Bayes IM 16 and Bayes C did not identify this QTL. The BTA 20 QTL is segregating with Angus, Red Angus, Hereford, and Simmental and ideally we want to be able to detect it.

Figure 4.3.3: YG: Genetic Variance for BTA 20 between 0 and 9 MB



a. The shaded region indicates the area with which we expect to observe peaks in the genetic variance.

Again, we are only going to focus on one of the three QTLs. The details for the other two QTLs can be found in Appendix A. The QTL we will focus on is BTA 20 at 4 MB. As mentioned above this QTL is segregating within 4 breeds, including our three major breed groups, Angus, Hereford, and Simmental. Saatchi et al. [52] identified an additional QTL for YG at 5 MB segregating in the Shorthorn and Simmental breeds. This second QTL at 5 MB is the QTL which was detected by the models that were able to detect a QTL.

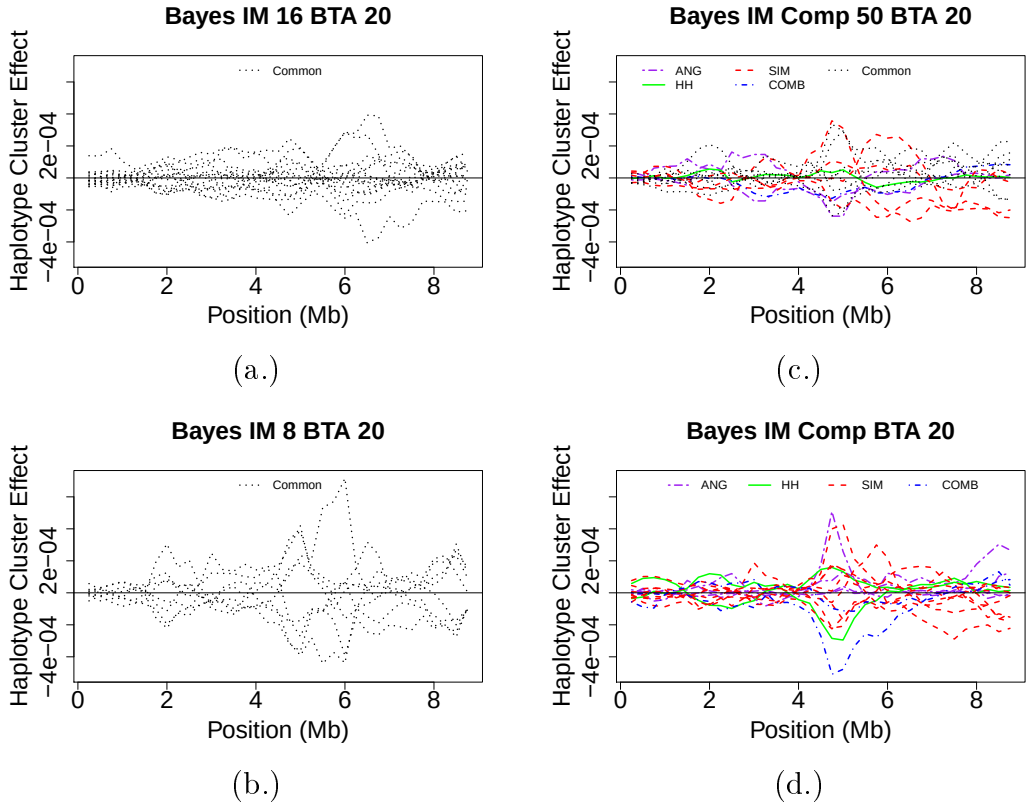
Figure 4.3.3 (a.) and (b.) shows the genetic variances for the individual

SNPs in Bayes B and Bayes C, respectively. Figure 4.3.3 (c.) through (f.) shows the genetic variances for the individual putative QTLs for the Bayes IM Models. Bayes B is showing several SNPs between 4 and 6 MB which are having a small effect. The effects observed within Bayes B are again larger than those observed within Bayes C due to the thicker tails of the t-distribution used by the Bayes B model. It should also be noted that Bayes B appears to be detecting a QTL right at 2 MB, which was also picked up within the Bayes C model. Only Bayes IM 8 from the Bayes IM models is detecting any increase in the genetic variance at 2 MB.

Bayes IM Comp and Bayes IM 8, which identified this QTL, are showing the largest peaks. Bayes IM Comp appears to have a clearly defined peak, whereas Bayes IM 8 is showing two peaks. Bayes IM Comp 50, which nearly identified this QTL, is showing a clearly defined peak that is similar to the peak observed within the Bayes IM Comp model, but the peak is not as large as the peak observed within the Bayes IM Comp model. Bayes IM 16 looks as if it is identifying the QTL at 7 MB rather than at 4 or 5 MB. Again, because the SNPs and the putative QTLs are on different scales, we cannot directly compare the Bayes B and Bayes C plots to the Bayes IM plots.

Moving to the haplotype cluster effects, which can be seen in Figure 4.3.4, we can observe that Bayes IM 16 is showing zero haplotype clusters with an effect at either 4 or 5 MB. There appear to be two haplotype clusters from Bayes IM 16 with a positive effect and one haplotype cluster with a negative effect around 7 MB which explains the peak we saw at 7 MB in Figure 4.3.3 (c.). Bayes IM 8 is showing one haplotype cluster with a large positive effect and three haplotype clusters with a negative effect, although the plot appears to be quite noisy. Bayes IM Comp 50 also appears to be quite noisy with no truly

Figure 4.3.4: YG: Haplotype Effect Estimates for BTA 20 Between 0 and 9 MB



large effect haplotype clusters. Bayes IM Comp 50 does have several medium effect haplotype clusters including two Simmental and two common haplotype clusters with a positive effect and one Angus, one common, and two Simmental with a negative effect. One of the Simmental haplotype clusters has a medium positive effect at 5 MB and a medium negative effect at 6 MB. The cross-over observed within Bayes IM Comp 50 may be due to the nature of a HMM. One explanations is that the model hit a local maximum and it was unable to do a good job of estimating the haplotype cluster. Bayes IM Comp has the clearest haplotype effect estimates. Bayes IM Comp has one large positive Angus and one large positive Simmental haplotype cluster. Additionally, there is one large negative combined breed group haplotype cluster and one Hereford haplotype

cluster with a medium effect. Bayes IM Comp appears to do the best job at detecting the breed specific QTL effects. However, since we do not know what the true QTL effects should be, it is better to determine the best model based on predictability.

4.3.2.3 WWT

Now we will examine the QTLs for the first weight trait, WWT. Direct weaning weight is the calf's weight at weaning due solely to growth. There were six identified QTLs for this trait and Table 4.3.5 identifies which of the QTLs were identified by the six models. The QTL on BTA 2 at 6 MB was only identified by Bayes IM 16 and was nearly identified by Bayes IM 8. The QTL on BTA 5 at 106 MB was nearly identified by Bayes C at 107 MB but was not identified by the other five models. The QTLs on BTA 6 at 38 MB, BTA 7 at 93 MB, BTA 14 at 24 MB, and BTA 20 at 4 MB were identified by all six models. The QTL on BTA 6 was one of the top two QTLs for every model except for Bayes B where it ranked 16th. Additionally, Bayes B and Bayes C identified the QTL on BTA 7 at 92 MB, while all four Bayes IM models identified this QTL as occurring at 93 MB. The QTL on BTA 14 was identified at 24 MB for all models except for Bayes IM Comp which identified this QTL at 25 MB. The four QTL which were identified well are also QTL which are segregating

Table 4.3.5: WWT: QTLs Identified in the top 100 1 MB Windows

BTA_MB	Bayes B	Bayes C	Bayes IM 16	Bayes IM 8	Bayes IM Comp 50	Bayes IM Comp
2_6	No	No	Yes (92)	No (185)	No	No
5_106	No	No** (193)	No	No	No	No
6_38	Yes (16)	Yes (2)	Yes (2)	Yes (2)	Yes (1)	Yes (2)
7_93	Yes* (26)	Yes* (23)	Yes (5)	Yes (4)	Yes (8)	Yes (4)
14_24	Yes (13)	Yes (13)	Yes (15)	Yes (8)	Yes (4)	Yes** (6)
20_4	Yes (7)	Yes (9)	Yes (6)	Yes (6)	Yes (5)	Yes (5)

a. * MB below, **MB above

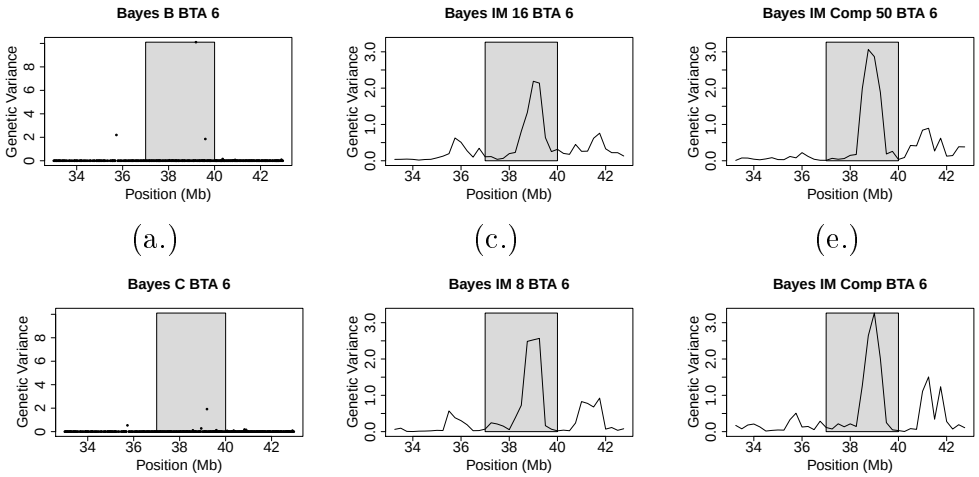
within the Simmental Breed.

As we did before, we are going to focus on one location, BTA 6 at 38 MB. The details for the other five locations can be found in Appendix A. The QTL on BTA 6 at 38 MB is segregating within the Gelbvieh, Hereford, Limousin, Red Angus, and Simmental breeds. It should be noted that Saatchi et al. [52] identified additional QTLs for WWT at 37 MB segregating within the Red Angus and Simmental Breeds, at 39 MB segregating within Shorthorn, Red Angus, and Simmental, and slightly further away at 41 MB which is segregating only within Shorthorn.

Using Figure 4.3.5 (a.) and (b.), we can examine the estimated genetic variance for each SNP within Bayes B and Bayes C. The genetic variance for each putative QTL within the Bayes IM models is in Figure 4.3.5 (c.) through (f.). Due to the thick tails of the t-distribution, Bayes B is detecting a much larger peak than Bayes C and has a SNP with a very large effect around 39 MB. Bayes C shows an elevation from the same SNP detected within Bayes B but puts this SNPs effect on a smaller scale. The four Bayes IM models appear to do equally well in detecting this QTL. Although the peak observed within the two Bayes IM Comp models is slightly taller than the peaks within the two Bayes IM models. In addition, the Bayes IM models seem to slightly detect the QTL at 41 MB and the Bayes B and Bayes C models show very little elevation around 41 MB.

We will now focus on the haplotype cluster effects estimated by the four Bayes IM models, which are presented in Figure 4.3.6. Bayes IM 16 and Bayes IM 8 are showing that every haplotype cluster has an effect. Approximately half of the haplotype clusters have a large or medium sized positive effect and half of the haplotype clusters have a large or medium sized negative effect. The

Figure 4.3.5: WWT: Genetic Variance for BTA 6 between 33 and 43 MB

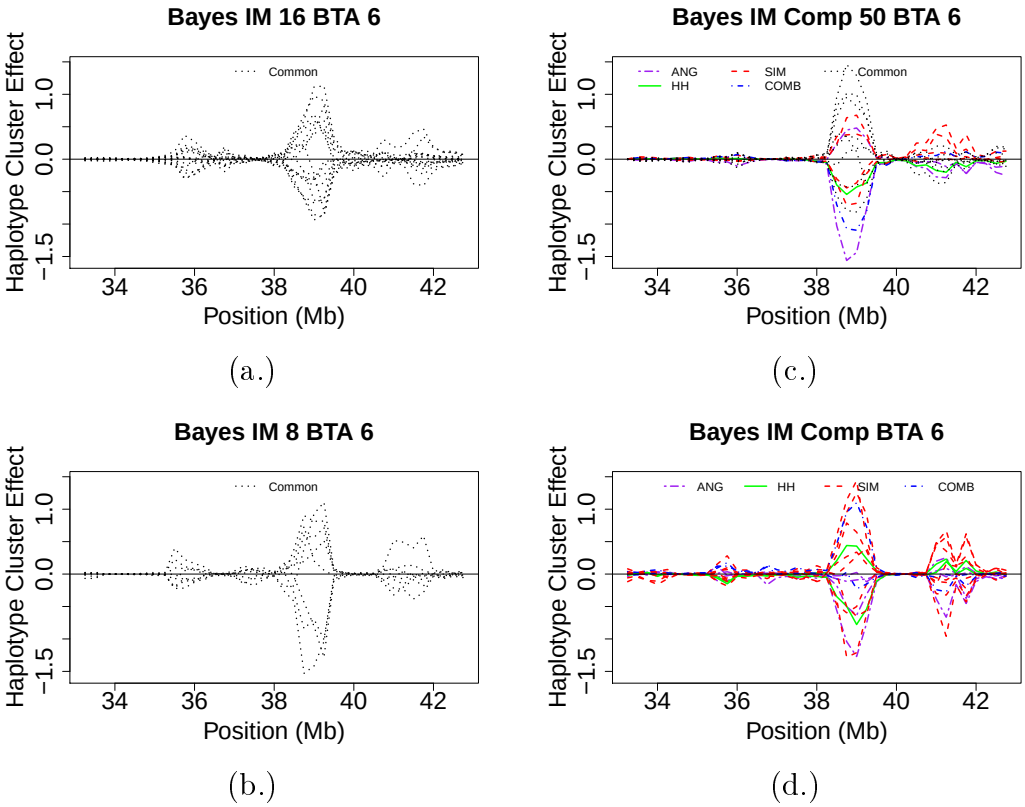


(a.) (c.) (e.)

(b.) (d.) (f.)

a. The shaded region indicates the area with which we expect to observe peaks in the genetic variance.

Figure 4.3.6: WWT: Haplotype Effect Estimates for BTA 6 Between 33 and 43 MB



(a.) (c.)

(b.) (d.)

haplotype clusters within the Bayes IM 8 model have a larger effect because there are a fewer overall number of cluster. This means the variance is being spread among only 8 haplotype clusters rather than 16 haplotype clusters in the case of Bayes IM 16.

Bayes IM Comp shows that about half of the Simmental haplotype clusters have a positive effect and the other half have a negative effect. The Hereford haplotype clusters are also split with one having a positive effect and one having a negative effect. One of the combined breed group haplotype clusters has a positive effect while the other cluster appears to have no effect. Finally, there are 2 Angus haplotype clusters with a negative effect and the other two clusters have no effect at all. Most of the positive Simmental, the one positive Hereford, and the positive combined breed group clusters appear to have been replaced by three common haplotype clusters within the Bayes IM Comp 50 model. Bayes IM Comp 50 is still detecting the negative Hereford and one of the negative Angus haplotype clusters, but now there is an Angus cluster with a medium positive effect and the combined breed group cluster has a large negative effect which is the opposite of what was occurring within Bayes IM Comp.

Ideally, the addition of the common haplotype clusters would combine the haplotypes which are identical across the different breeds which in turn would allow us to better see the haplotype effects which are unique to a specific breed and better estimate those effects. We examined common haplotype cluster 11, which is the common haplotype cluster with the largest effect within Bayes IM Comp 50. We then calculated the probability than an individual is in a particular cluster within Bayes IM Comp given they are in haplotype cluster 11 within Bayes IM Comp 50. This revealed that there are five Simmental

haplotype clusters within Bayes IM Comp that have a strong mapping. Given an individual is in common haplotype cluster 11 within Bayes IM Comp 50, the probability they are in Bayes IM Comp Simmental haplotype cluster 9 is 9%, Simmental cluster 10 is 25%, Simmental cluster 12 is 8%, Simmental cluster 13 is 23%, and Simmental cluster 15 is 19%. It should also be noted that these five haplotype cluster all have a positive effect within Bayes IM Comp. This explains why common haplotype 11 has such a large effect in Bayes IM Comp 50. All of the positive effects from Bayes IM Comp are being combined together into one haplotype cluster. Additionally, we would have liked to observed haplotype clusters from other breeds mapping to this common haplotype cluster, which did not occur here.

4.3.2.4 YWT

The last trait we will examine for QTL identification is the weight trait YWT. This refers the weight of the calf taken between 320 and 440 days of age, or when the calf is approximately a year old. There were five identified QTLs for YWT, these were identical to the QTLs identified for WWT except that the QTL on BTA 2 is not present for YWT. By examining Table 4.3.6 we see similar patterns observed for YWT and WWT. The QTL on BTA 5 at 106 MB was nearly identified by Bayes IM Comp, but was not identified by any other model. The other four QTL were identified by all six models. Again,

Table 4.3.6: YWT: QTLs Identified in the top 100 1 MB Windows

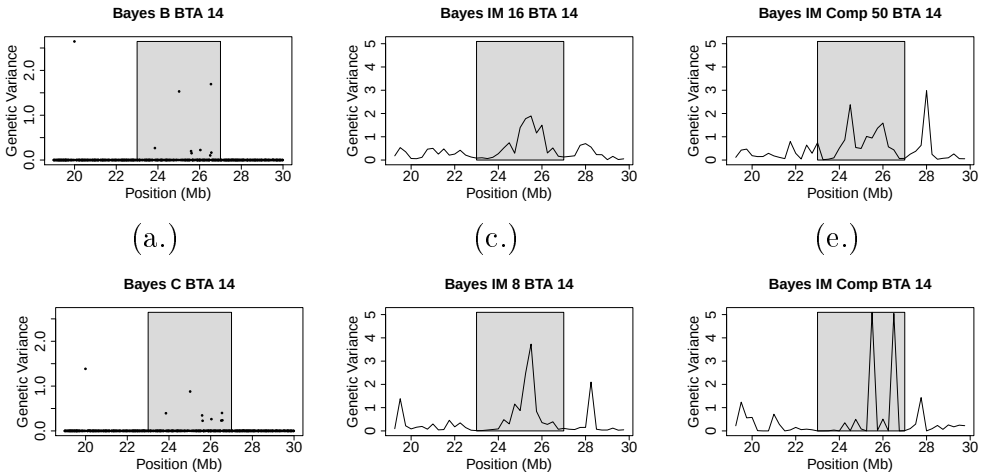
BTA_MB	Bayes B	Bayes C	Bayes IM 16	Bayes IM 8	Bayes IM Comp 50	Bayes IM Comp
5_106	No	No	No	No	No	No (118)
6_38	Yes (6)	Yes (2)	Yes (2)	Yes (2)	Yes (2)	Yes (2)
7_93	Yes* (25)	Yes* (28)	Yes (8)	Yes (5)	Yes (13)	Yes (5)
14_24	Yes (32)	Yes (32)	Yes (22)	Yes (30)	Yes (6)	Yes** (8)
20_4	Yes (3)	Yes (5)	Yes (5)	Yes (6)	Yes (4)	Yes (4)

a. * MB below, **MB above

Bayes B and Bayes C identified the QTL on BTA 7 at 92 MB rather than 93 MB and Bayes IM Comp identified the QTL on BTA 14 at 25 MB, rather than 24 MB. Both the QTL on BTA 6 and the QTL on BTA 20 were identified as a top 10 QTL by all six models.

Since we have examined the QTLs on BTA 6, 7, and 20 above and the haplotype effect estimates show similar patterns, we are going to focus on the QTL on BTA 14 here. Although, the details for the other four QTL can be found in Appendix A. The QTL on BTA 14 at 24 MB is segregating within the Gelbvieh and Simmental breeds. Additional QTLs for BTA 14 identified by Saatchi et al. [52] include a QTL at 23 MB which is Simmental specific, a QTL at 25 MB which is segregating within Gelbvieh and Simmental, and a QTL at 26 MB which is segregating within Brangus and Simmental.

Figure 4.3.7: YWT: Genetic Variance for BTA 14 between 19 and 30 MB



a. The shaded region indicates the area with which we expect to observe peaks in the genetic variance.

The estimated individual SNP genetic variance for Bayes B and Bayes C are shown in Figure 4.3.7 (a.) and (b.) and the estimated putative QTL genetic variances for the Bayes IM models are shown in Figure 4.3.7 (c.) through

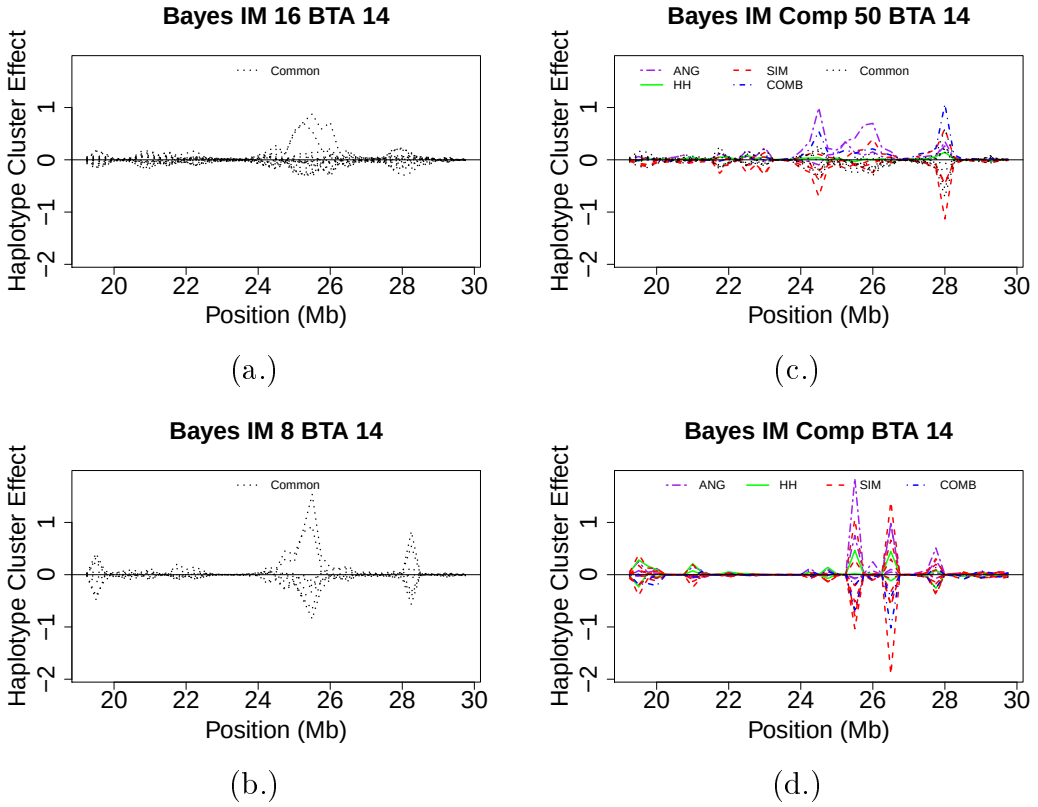
(f.). Bayes B and Bayes C are showing several SNPs with elevated effects. As we have observed before, Bayes B has much larger SNP effect estimates than Bayes C due to the thicker tails in the t-distribution.

For Bayes IM Comp and Bayes IM Comp 50, this QTL ranked in the top 10. Bayes IM Comp 50 is showing two separate peaks one within the 24 MB window and one within the 25 MB window. Bayes IM Comp is also showing two peaks which are much taller than those observed within Bayes IM Comp 50 and they have both been shifted to the right by one megabase. Bayes IM 8 and Bayes IM 16, on the other hand, are showing one clearly defined peak within the range of 24 to 27 MB. The peak on Bayes IM 8 is larger than the peaks seen for Bayes IM 16 and Bayes IM Comp 50, but still smaller than the peak seen within Bayes IM Comp.

The final step is to observe the patterns seen within the haplotype cluster effects for the Bayes IM models. Figure 4.3.8 displays these plots. Recall that in Figure 4.3.7, Bayes IM 8 and Bayes IM Comp had much larger effect sizes than Bayes IM 16 and Bayes IM Comp 50. This trend can also be observed when looking at the haplotype cluster effects.

Bayes IM 16 is only showing two haplotype clusters with a positive effect while Bayes IM 8 is showing two clusters with a positive effect and two clusters with a medium sized negative effect. These negative effect clusters were not present within the Bayes IM 16 because the effects from these negative haplotype clusters in Bayes IM 8 are being spread among several clusters within Bayes IM 16 causing the effect to be muted. For example, we calculated the probability of being in a particular haplotype cluster within Bayes IM 16 given we were in haplotype cluster 2 within Bayes IM 8, which is one of the negative haplotype clusters. This revealed that the probability was 12% of being in

Figure 4.3.8: YWT: Haplotype Effect Estimates for BTA 14 Between 19 and 30 MB



haplotype cluster 3, 9% of being in haplotype cluster 5, 7% of being in haplotype cluster 7, and 29% of being in haplotype cluster 9 within Bayes IM 16. Thus the effects for Bayes IM 8 cluster 2 is being spread among all four of the clusters from Bayes IM 16.

Bayes IM Comp is showing two clearly defined QTLs. One in the 25 MB window and one in the 26 MB window. The 25 MB window is showing an Angus haplotype cluster with a large positive effect along with medium sized effects from a Simmental, Hereford, and additional Angus haplotype cluster. The negative side is showing three Simmental and one combined breed group haplotype clusters with medium sized effects. Within the 26 MB window, the large positive Angus effect was been reduced and both the large positive and

large negative Simmental haplotype cluster effects have increased. The large effects that go to zero and get large again within the Bayes IM Comp model may be due to the nature of an HMM and this is a location where our haplotype clusters are not being estimated well. Bayes IM Comp 50 is not showing two clearly separated peaks like what was observed within Bayes IM Comp, but we can still observe two peaks, one within the 24 MB window and one within the 25 MB window. At 24 MB there is an Angus haplotype cluster with a large positive effect and a combined breed group haplotype cluster with a medium sized effect. There are also several Simmental haplotype clusters with medium sized negative effects. At 25 MB the only cluster with a significant effect is the one Angus haplotype cluster that also had a positive effect at 24 MB.

We observed two general trends throughout the investigation of the QTLs. The first trend is that, in general, the Bayes IM Comp 50 model ranks the known QTLs higher than any other model. This is an indication that this model is doing the best job at detecting the QTLs. The second trend is, for the haplotype effects themselves, Bayes IM Comp 50 tended to have less noisy haplotype estimates with QTL peaks which were clearly identified.

4.3.3 Prediction Accuracy

Traditionally, the best model is chosen based on which model has the highest prediction accuracy. Prediction accuracy was measured using a bivariate individual model which included the estimated breeding value of genotyped individuals and the weighted deregressed EPD to estimate the genetic correlations using ASReml v4.1 software [18]. The model for the estimated breeding value included a fixed intercept effect, random additive genetic effect, and a residual with a fixed variance of 0.0001% of the unweighted phenotypic vari-

ance of the deregressed EPD. The model for the deregressed EPD included a fixed intercept effect, random additive genetic effect, and a weighted residual with variance equal to $\frac{\sigma_e^2}{w_i}$, where $\frac{1}{w_i}$ is the weights according to the reliabilities of an individuals deregressed EPDs and are the same values which were used in training for the estimation of the marker effects. Additionally, the additive genetic effect and unweighted residual variances were fixed at 40% and 60%, respectively, of the deregressed unweighted phenotypic variance of the EPD. Each fold served as the evaluation set once and genetic correlations were predicted based on the pooled marker effects from using all folds except the evaluation fold.

Table 4.3.7: Prediction Accuracy for Low Simmental Fold

Traits	Bayes B	Bayes C	Bayes IM 16	Bayes IM 8	Bayes IM Comp 50	Bayes IM Comp
BWT	0.647 (0.042)	0.596 (0.044)	0.596 (0.043)	0.564 (0.045)	0.610 (0.042)	**0.612 (0.043)**
MILK	0.124 (0.077)	0.138 (0.077)	0.191 (0.077)	0.155 (0.077)	0.179 (0.078)	**0.195 (0.078)**
MWWT	0.464 (0.066)	0.468 (0.067)	**0.477 (0.067)**	0.431 (0.067)	0.465 (0.067)	0.436 (0.068)
WWT	0.561 (0.052)	0.546 (0.053)	0.568 (0.053)	**0.579 (0.053)**	0.573 (0.053)	0.572 (0.052)
YWT	0.595 (0.053)	0.583 (0.053)	0.595 (0.053)	**0.616 (0.053)**	0.575 (0.054)	0.563 (0.054)
CWT	0.683 (0.052)	0.659 (0.052)	0.667 (0.053)	**0.711 (0.052)**	0.692 (0.052)	0.648 (0.054)
BFAT	0.326 (0.051)	0.356 (0.051)	0.330 (0.064)	0.319 (0.065)	0.301 (0.065)	**0.352 (0.064)**
MARB	0.357 (0.084)	0.376 (0.083)	0.436 (0.081)	0.439 (0.081)	0.422 (0.082)	**0.440 (0.081)**
REA	0.454 (0.089)	0.451 (0.088)	0.459 (0.088)	0.441 (0.088)	**0.477 (0.088)**	0.445 (0.089)
YG	0.462 (0.081)	0.455 (0.082)	**0.397 (0.084)**	0.381 (0.085)	0.379 (0.084)	0.378 (0.085)
CE	0.680 (0.044)	0.669 (0.045)	0.662 (0.044)	0.663 (0.044)	0.668 (0.044)	**0.678 (0.043)**
DOC	0.523 (0.095)	0.501 (0.094)	**0.581 (0.090)**	0.426 (0.096)	0.514 (0.093)	0.526 (0.092)
MCE	0.425 (0.069)	0.431 (0.069)	**0.438 (0.068)**	0.369 (0.070)	0.432 (0.068)	0.402 (0.068)

a. Genetic correlation (SE)
b. Best model out of all 6 is underlined
c. Best model out of all Bayes IM models is in bold.

The prediction accuracy of the predictors trained in the high and medium Simmental folds and evaluated in the low Simmental fold provides a measure of how well each of the models worked when predicting the genetic merit of individuals whose breed composition differed from the the majority of individuals in training. The results for the low Simmental fold is presented in Table 4.3.7 Results for the high Simmental and medium Simmental folds can be found in

Appendix A Tables A.3.1 and A.3.3. In a homogeneous population, Bayes B and Bayes C tend have a higher prediction accuracy than the original Bayes IM model [28]. The fact that Bayes B is only the best overall model for three traits and Bayes C for one trait shows great promise for the Bayes IM models.

There are several questions which we hoped to answer. We wanted to know whether doubling the number of haplotype clusters from 8 to 16 common haplotype clusters increased our prediction accuracy (Bayes IM 16 versus Bayes IM 8). We wanted to know if the addition of the common haplotype clusters into the Bayes IM Comp model improved our prediction accuracy (Bayes IM Comp versus Bayes IM Comp 50). Additionally, we wanted to know whether the inclusion of breed specific haplotype clusters improved our prediction ability (Bayes IM vs Bayes IM Comp). In order to answer these questions, we looked at the the score differential for each model. The score differential was calculated by taking the number a traits out of 13 that model A had a high prediction accuracy than model B minus the number of traits model A had a lower prediction accuracy than model B. A positive score differential implies model A had a higher overall prediction accuracy than model B and a negative score differential implies model A had a lower overall prediction accuracy than model B. The score differentials for the low Simmental fold are presented in Table 4.3.8. The high and medium fold score differential tables can be found in Appendix A Tables A.3.2 and A.3.4.

For the comparison between Bayes IM 16 and Bayes IM 8, Bayes IM 16 had better prediction for 8 traits, Bayes IM 8 had better prediction for 4 traits, and there was a tie between to two models for one trait. This suggests that doubling the haplotype clusters from 8 to 16 had a positive effect on the prediction accuracy. For Bayes IM Comp versus Bayes IM Comp 50, Bayes

Table 4.3.8: Score Differential for the Low Simmental Fold

	Model A					
Model B	Bayes B	Bayes C	Bayes IM 16**	Bayes IM 8	Bayes IM Comp 50	Bayes IM Comp
Bayes B		-3	4	-3	0	-3
Bayes C	3		6	-3	3	-1
Bayes IM 16	-4	-6		-4	-3	-1
Bayes IM 8	3	3	4		1	5
Bayes IM Comp 50	0	-3	3	-1		1
Bayes IM Comp	3	1	1	-5	-1	
Total Score Differential	5	-8	18**	-16	0	1

a. Score differential = (# of traits Model A had higher prediction accuracy) - (# of trait Model B had higher prediction accuracy)
b. ** represents the model with the highest total score differential

IM Comp had better prediction for 6 traits, Bayes IM Comp 50 had better prediction for 5 traits and for 2 traits there was a tie, suggesting the addition of the common haplotype clusters had little effect on the prediction accuracy. For the two Bayes IM models versus the two Bayes IM Comp models, the Bayes IM 16 model had better prediction than either Bayes IM Comp model, suggesting there was little benefit to using the breed specific haplotype clusters rather than using a set of common haplotype clusters. This makes sense as the low fold consisted of individuals whose breed composition was mainly composed of breeds other than the Simmental breed. The Bayes IM Comp models do not contain enough Angus, Hereford, and combined breed group haplotype clusters to accurately capture the true genetic architecture of this group of individuals. Additionally, the added computational cost of the Bayes IM Comp models was of little benefit.

Additionally, we broke the traits up into weight traits, carcass traits, and

other traits and once again there was no clear pattern which determined a superior model based on the type of trait. We also examined across all three folds to determine if, for each trait, the same model was superior for all three folds and this did not hold. From a prediction standpoint, all six models performed similarly and this is consistent with other studies comparing the prediction accuracy of different models.

4.3.4 Conclusions

Overall, there are several lessons to be learned from this data set. When it comes to QTL identification, the Bayes IM models tend to detect the QTLs much clearer than the Bayes B and Bayes C models. Additionally, Bayes IM Comp tends to produce the largest peaks for the QTLs, which makes since as the Bayes IM Comp models have a larger overall genetic variance. The advantage of the Bayes IM Comp models is that all haplotype clusters are breed specific so we are able to identify which breeds the QTL is present in for a given location.

We mentioned above that occasionally we see haplotype cluster estimates that go from being a large positive effect to immediately being a large negative effect or vice versa, which may be due to the model doing a poor job of estimating the clusters. There are a several options to address this issue. One possibility is to fit more than one HMM model and then perform model averaging over all the HMMs. This is the route that fastPHASE uses [56]. A second option involves the use of phased genotypes. Currently, we are using unphased genotypes to train our HMM. There are several methods available that can phase our genotypes. Using phased genotypes may allow us to better estimate the haplotype clusters. A third option is to use more purebred indi-

viduals in the training of the HMM. There are only 311 purebred Simmental and 32 purebred Angus individuals, where a purebred individual refers to any individual whose breed composition for a given breed is greater than $\frac{7}{8}$. It is possible that the model is having a hard time determining which haplotype clusters belong to which breed. By including purebred individuals from each breed into the training of the HMM, we should be able to better estimate the haplotype clusters. This technique was used for the simulated data set below and the reproductive longevity data set in Chapter 5 and it allowed us to better define the haplotype clusters.

4.4 Evaluation of Bayes IM Comp Using a Simulated Data Set

4.4.1 Population Structure

Simulated data gives us the advantage of being able to know where the true QTLs are. To better access the models performance in identifying QTLs, a bovine population was simulated using the QMSim software v1.10 [54]. Simulation parameters where based on a simulation preformed by Brito et al. [6]. The population was simulated using a forward-in-time process [7] with 172,532 SNPs, 61,255 of which were polymorphic, and 1000 QTLs, 361 of which were polymorphic, randomly placed across 29 autosomes. The phenotypic trait was generated with a heritability of 0.4 and phenotypic variance of 1. Breeding values were then estimated using BLUP, which is based on Henderson's mixed model linear equations [24]. The first phase of the simulation creates a historical population. This was done in three steps. First, 1000 generations with a constant size of 1000 were generated. Second, 1000 generations with a gradual population decrease from 1000 to 200 were generated. Finally, 20 more generations with a constant population size of 200 were generated. This process

creates initial linkage disequilibrium and establishes mutation-drift equilibrium among the historical population. For all three steps, the number of individuals of each sex remained equal and the mating system was a random union of gametes.

In the second phase, we created two unique lines, a high and a low line. For both lines, we selected 50 male and 50 female founders from the last generation of the historical population. The 50 males and 50 females with the highest phenotype were used in the high line and the other 50 males and females were used in the low line. In order to expand the population, 20 generations were generated where each dam produced a litter of size 10 with an equal number of male and female offspring. The mating design was again random union of gametes and selection was preformed based on phenotype. For the low line, we selected based on low phenotypes and for the high line, we selected based on high phenotypes.

The final phase of the simulation was to generate the cross-breed population. The founders consisted of 250 males from the last generation of both the low and high lines and 1250 females from the last generation of both the low and high lines. A total of 5 generations were generated using a random mating design with random selection and culling based on age. Each dam had a litter size of one and the number of male and female offspring were equal. The sire replacement rate was 0.6 and a dam replacement rate was 0.2. A summary of how the populations were generated can be found in Appendix B, Figure B.1.1.

4.4.2 Genome

In order to better mimic the true architecture of the bovine genome, the length of the 29 autosomes were based on the Btau_3.1 assembling [59] and had a total length of 2,333 megabases. The SNP loci were randomly placed along each chromosome. The initial number of SNPs per chromosome was chosen based on the BovineSNP50 v3 BeadChip by Illumina [25]. We counted number of SNPs per chromosome on the BovineSNP50 chip and multiplied this number by 3.33. This ensured that we had approximated 60,000 polymorphic SNPs in the final analysis. The total number of QTL was chosen to be 1000 since only approximately one-third of all QTL will be active and we were looking to simulate approximately 300 active QTL. First, each of the 1000 QTL were randomly assigned to one of the 29 automosomes and then they were randomly distributed across their respective chromosomes. Following Brito et al. [6], the mutation rate for the SNPs and the QTLs was 10^{-5} and the additive allelic effects were sampled from a gamma distribution with shape parameter equal to 0.4. Both the rate of missing marker genotypes and the rate of marker genotyping error were set to be 0. A summary of the number of simulated SNP markers and QTLs for each autosome is given in Appendix B, Table B.1.1.

4.4.3 Models Compared

Five models were considered: Bayes B, Bayes C, Bayes IM (100% common, this is the Bayes IM model with a total of 16 haplotype clusters shared by both lines), Bayes IM Comp 50 (50% common, this is the Bayes IM Comp model with 4 haplotype clusters assigned to the low line, 4 assigned to the high line, and 8 assigned to be common among both lines for a total of 16

haplotype clusters), and Bayes IM Comp (0% common, this is the Bayes IM Comp model with 8 haplotype clusters assigned to the low line, 8 haplotype clusters to the high line, and zero haplotype clusters assigned to be common among both lines for a total of 16 haplotype clusters).

In the historical population, the phenotypes were generated with a variance of 1 and a heritability of 0.4. This assumes that environmental or residual variance is equal to 0.6 and genetic variance is equal to 0.4. As time passed, the genetic variance slowly increased which increased the phenotypic variance. By time we created our crossbred population, the phenotypic variance was approximately 15. In order to get back to a phenotypic variance of 1 and genetic variance of 0.4, we scaled each individuals genetic value such that the genetic variance was equal to 0.4. In addition to the SNP effects, the model included an overall mean, an effect for generation, and an effect for sex as fixed effects. The residual for individual i was assumed to be $N(0, \sigma_e^2)$.

4.4.4 Training and Evaluation Sets

Initially the HMM was trained using the individuals from generation five of the cross breed population. This led to poor haplotype cluster assignments since the model was unable to accurately determine which haplotypes belonged to which lines since generation five was the most diverse from a composition standpoint. In order to overcome this, the 3,000 parents which founded the cross breed population were utilized in a two-step process to estimate the parameters for the HMM. In the first phase, we used only the 3,000 parents in order estimate the haplotype clusters. In the second phase, we added the 2,500 individuals from generation five of the cross breed population to the parents in order to estimate the transition parameters r_B and r_W for Bayes IM Comp

and r for Bayes IM.

Once the parameters were estimated for the HMM, all five cross breed population generations were used to create the training and evaluation sets. Similar to the Simmental data set, three folds were created by partitioning the data into a fold of individuals with high phenotypes, a fold with medium phenotypes, and a fold with low phenotypes. The high phenotype fold contained individuals with an average breed composition of 85% from the high line and 15% from the low line. The medium fold contained individuals with an average breed composition of 50% from the high line and 50% from the low line. The low fold contained individuals with an average breed composition of 15% from the high line and 85% from the low line. There were 3,526 individuals in the high fold, 5,548 in the medium fold, and 3,426 in the low fold for a total of 12,500 individuals. In each model above, two folds were used for training and the third fold was used for evaluation to determine prediction accuracy. QTL detection used all three fold together.

4.5 Results and Discussion for the Simulated Data Set

4.5.1 Posterior Distribution Comparison

The statistical definition of heritability is the proportion of phenotypic variance explained by the genetic variance. Since this was simulated data with a known heritability of 0.4, the prior for the genetic variance was estimated to be equal to 40% of the variance seen in the phenotypes, which was equal to 1. The remaining 60% of the variance was assigned to the prior for the residual

variance. The haplotype effect variance is estimated to be:

$$\sigma_b^2 = \frac{\sigma_g^2}{\left((1-\pi)\cdot n_{QTL}\cdot\left(1-\frac{1}{N_{cluster}}\right)\cdot 2\right)},$$

where σ_g^2 $(= 0.4)$ is the genetic variance, π $(= 0.975)$ is the proportion of markers which have no effect, n_{QTL} $(\approx 10,000)$ is the number of putative QTL in the model, and $N_{cluster}$ $(= 16)$ is the number of haplotype clusters in the model. Therefore, the prior for the haplotype effect variance was originally estimated to be 0.0008, but this was shown to be much too high and a new prior of 0.0003 was used. Table 4.5.1 summarizes the priors, posterior means, and standard errors for each of the five models evaluated. Since Bayes B and Bayes C are not haplotype models they do not have a haplotype effect variance.

Table 4.5.1: Prior and Posterior Means (SE) for Variance Components

Model	Genetic Variance	Residual Variance	Heritability	Haplotype Effect Variance
Prior	0.4	0.6	0.4	0.0003
Bayes B	0.425 (0.009)	0.568 (0.008)	0.428 (0.006)	N/A
Bayes C	0.406 (0.009)	0.571 (0.008)	0.415 (0.006)	N/A
Bayes IM	0.443 (0.010)	0.533 (0.008)	0.454 (0.008)	0.00025 (0.00002)
Bayes IM Comp 50	0.496 (0.013)	0.480 (0.011)	0.508 (0.011)	0.00033 (0.00003)
Bayes IM Comp	0.539 (0.019)	0.438 (0.019)	0.552 (0.019)	0.00040 (0.00004)

The residual variance prior is higher than the posterior mean for all five models, which is consistent to what was observed within the Simmental data set. The Bayes B and Bayes C models are estimating the residual variance to be slightly higher than any of the Bayes IM versions. Again, Bayes IM Comp is estimating the lowest posterior residual variance.

Also similar to the Simmental data set, Bayes B and Bayes C have the lowest posterior means for the genetic variance and the Bayes IM Comp models have the highest posterior means. This is consistent with the trend observed within the residual variance since as the estimated residual variance decreases,

the estimated genetic variance will increase since the overall variance is being shared between the genetic and residual variances.

Overall, the posterior mean heritability estimates are greater than the 0.4 value. Although, the heritability values for the simulated data set are not as extreme for the Bayes IM Comp models as what was observed within the Simmental data set. The higher residual variance estimates and lower genetic variance estimates for the Bayes B and Bayes C models leads to the small heritability estimates. Conversely, the lower residual variance estimates and higher genetic variance estimates observed from the Bayes IM Comp models leads to the larger heritability estimates.

The posterior means for the Bayes B and Bayes C models are closest to the prior values and the Bayes IM Comp models are the farthest from the prior values. Again, one explanation for this is that the Bayes IM models are not influences as strongly by the prior information as Bayes B or Bayes C. A second possible explanation is that the haplotypes in the Bayes IM models are doing a better job of capturing the true genetic variance than the SNPs are in the Bayes B and Bayes C models.

Comparing the haplotype effect variances for the three Bayes IM models, we can observe the same trend as we observed with the genetic variance. Bayes IM has the smallest posterior mean and Bayes IM Comp has the highest posterior mean. This makes sense as the haplotype effect variance is a function of the genetic variance.

The posterior distribution plots, Figures B.2.1 and B.2.2 in Appendix B, were also examined for abnormalities. Overall, all the distribution plots were symmetric and bell shaped, which is an indication that the initial burn-in was large enough for the model to settle. The posterior distributions are not

centered at the prior used, however the prior was contained within the posterior distributions for all five models.

Finally, the trace plots were examined as a second check that the initial burn-in and overall number of MCMC samples was large enough. Overall, the trace plots for Bayes B, Bayes C, and Bayes IM appeared random with no obvious patterns. This is another indication that the initial burn-in is sufficient and a larger number of MCMC samples is not necessary in order to get a good estimate of the posterior mean for the parameters. The only issues were the trace plots for the residual variance for Bayes IM Comp and Bayes IM Comp 50. There is a downward trend in the residual variance values. This is an indication that a larger number of MCMC samples is needed in order for the model to be more stable.

4.5.2 QTL Identification and Haplotype effect estimates

The genetic variance of each polymorphic QTL is equal to $2pq\alpha^2$, where p is the frequency of the A allele in the composite population, q is the frequency of the B allele in the composite population, and α is the effect of A allele minus the effect of the B allele. The QTLs were then ranked from largest to smallest based on their genetic variance. The top five QTLs were identified and are summarized in Table 4.5.2. For each QTL, we included the A allele frequency for the high and low line parents of the composite population, which refers to generation 20 of the high and low line populations.

As we did with the Simmental data set, we broke the genome up into 1 MB windows and calculated the genetic variance within that window and ranked all the windows from largest to smallest. We looked at the windows which were 1 MB below, 1 MB above, and at the identified QTL and of these the

top rank was reported in Table 4.5.2. The ability of each model to identify QTLs was evaluated by comparing the rank of the genetic variance for each 1 MB window to the true rank of each QTL.

The Bayes B and Bayes C models rank the top 5 QTL regions within the top 20 windows. The Bayes IM and Bayes IM Comp models rank the top 5 QTL regions higher Bayes IM Comp 50. To compare how well each model does in identifying the location of the QTL, we looked at the QTL with the largest genetic variance (BTA 4 at 6.6 MB) and one of the ones with a moderate genetic variance (BTA 14 at 10.8 MB). The results for BTA 28 at 10.8 MB was similar to those for BTA 4 at 6.6 MB. The results for BTA 5 at 114 MB and BTA 1 at 113 MB was similar to those for BTA 14 at 10.8 MB.

Table 4.5.2: Top QTLs for the Simulated Data Set

			A Allele Freq.		Rank				
BTA	Pos.	GenVar	High Line	Low Line	Bayes B	Bayes C	Bayes IM	Bayes IM Comp 50	Bayes IM Comp
4	6.6	0.00086	0.002	0.998	10	14	16*	51	25
28	10.8	0.00084	0.994	0.003	14	18	52*	99	16**
5	114	0.00069	0.996	0.002	5*	4*	4	2*	4
14	10.8	0.00068	0.006	0.995	2	1	3	13	2
1	113	0.00054	0.005	0.980	8*	7*	41*	10*	5*

a. * MB below, **MB above

4.5.2.1 QTL on BTA 4

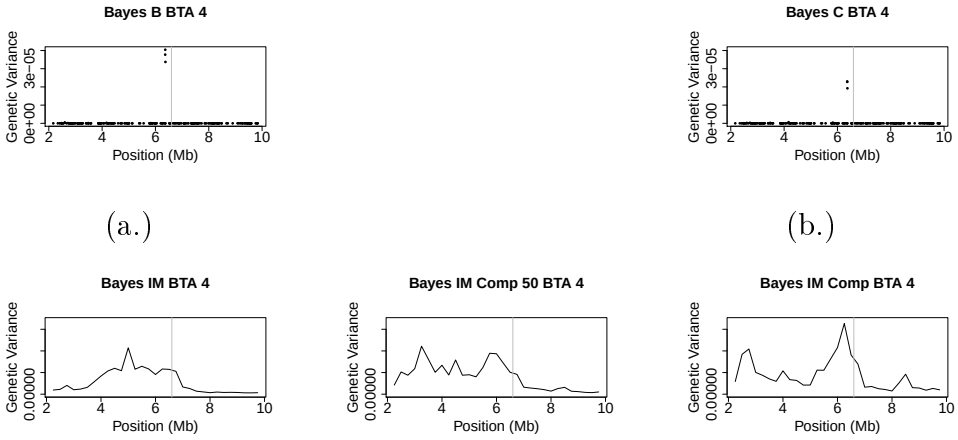
The top ranked QTL is on BTA 4 at 6.6 MB. Figure 4.5.1 shows the location of the true QTL compared to the SNP genetic variance for Bayes B and Bayes C and the putative QTL genetic variance for the Bayes IM models. Since the genetic variance is being spread across the SNPs or putative QTLs around the true QTL the peaks for each model should not be as large as the effect of the true QTL, reported in Table 4.5.2.

Based on Figures 4.5.1 (a.) and (b.), Bayes B is showing a larger peak in the SNP genetic variance than Bayes C. This result is similar to those observed

within the Simmental data set and is due to Bayes B using a t-distribution to estimate the SNP effects. The thicker tails of the t-distribution leads to larger SNP genetic variances. Bayes B and Bayes C identify the true location of the QTL and there is no additional noise.

For Figures 4.5.1 (c.), (d.), and (e.), the Bayes IM Comp model is showing the largest QTL effect at 6.6 of all the haplotype based models. However, it is identifying additional noise between 2 and 3 MB. Unlike Bayes B and Bayes C, Bayes IM and Bayes IM Comp 50 were unable to precisely locate the QTL exactly where the true QTL is since the elevation in genetic variance is spread across several megabases.

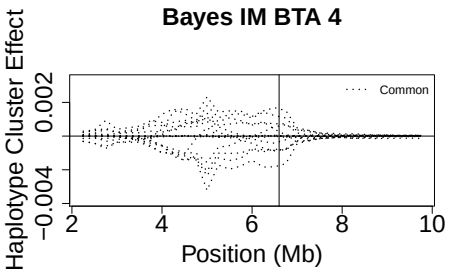
Figure 4.5.1: QTL Identification for BTA 4 Between 2 and 10 MB



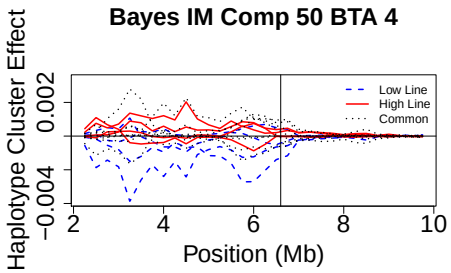
a. The vertical line indicates the location of the true QTL.

For this QTL, the A allele is having a negative effect, while the B allele has a positive effect. Since the A allele frequency for the high line was 0.002 and the A allele frequency for the low line was 0.998, we would expect to see the haplotype clusters from the high line having a positive effect and the haplotype clusters from the low line having a negative effect. Figure 4.5.2

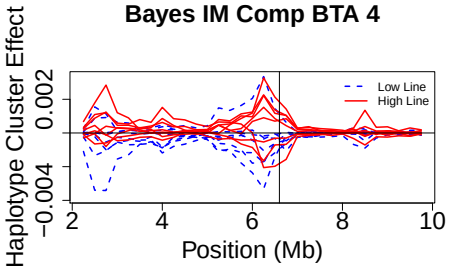
Figure 4.5.2: Haplotype Effect Estimates for BTA 4 Between 2 and 10 MB



(a.)



(b.)



(c.)

shows the haplotype cluster effects for the three Bayes IM models.

Bayes IM is showing four haplotype clusters are having a positive effect and four clusters are having a negative effect, the other eight haplotype clusters have no effect at all. Bayes IM is showing a clear peak for this QTL unlike above in Figure 4.5.1. There are still additional elevated haplotype effects between 2 and 4 MB which is still an indication that Bayes IM is unable to definitively identify the location of the true QTL.

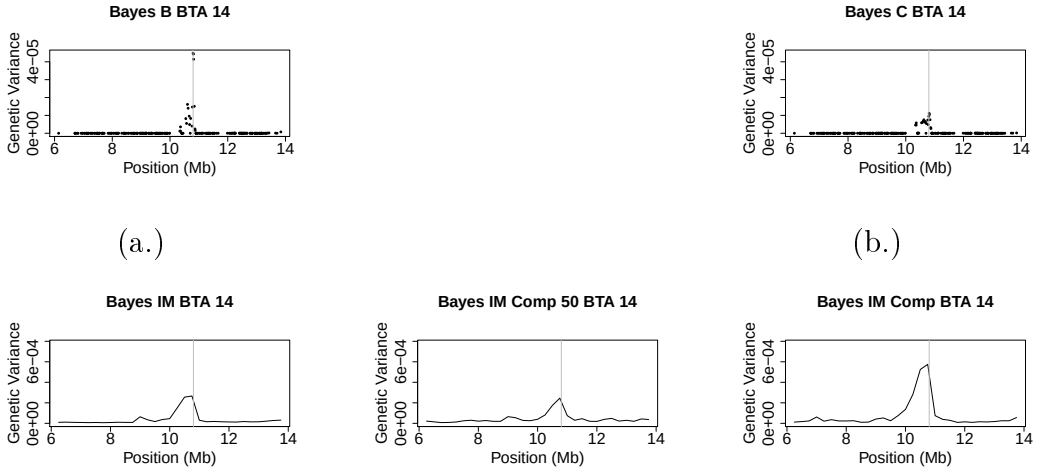
Bayes IM Comp is showing five high line haplotype clusters with a positive effect, two with a negative effect, and one with no effect. Additionally, there are four low line haplotype clusters with a negative effect, one with a very large positive effect, and two with no effect at all. The large positive low line

and the two negative high line clusters are slightly out of place, but since the A allele frequency is not fixed at one these are still possible. The QTL is better identified by this model, however there is still additional noise that is occurring around 2 MB.

At 6 MB, Bayes IM Comp 50 is showing two low line and one common haplotype cluster with a negative effect and three common haplotype clusters with a positive effect. The remaining clusters effects are close to zero. Overall, Bayes IM Comp 50 is very noisy and the true location of the QTL is not identified well.

4.5.2.2 QTL on BTA 14

Figure 4.5.3: QTL identification for BTA 14 between 6 and 14 MB



a. The vertical line indicates the location of the true QTL.

The QTL on BTA 14 at 10.8 MB is the fourth largest QTL. Every model is ranking this QTL in the top three except for Bayes IM Comp 50, which ranks this QTL as 13th. The true location of the QTL and the individual

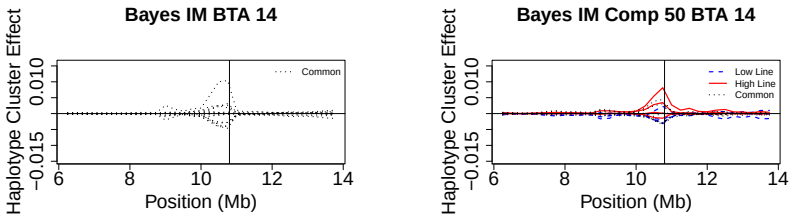
SNPs genetic variance for Bayes B and Bayes C or the putative QTLs genetic variance for the Bayes IM models is plotted in Figure 4.5.3.

Again, Bayes B is showing a much larger peak than the Bayes C model, which is consistent with Bayes B having thicker tails since it uses a t-distribution. Both Bayes B and Bayes C identify to true location of the QTL. Compared to the QTL on BTA 4, there are a larger number of SNPs here with a positive effect, which makes the QTL on BTA 14 clearer.

Unlike the QTL on BTA 4, all three Bayes IM models are able to clearly identify the QTL on BTA 14 at 10.8 MB. Bayes IM Comp has the largest peak out of the three Bayes IM models. Bayes IM and Bayes IM Comp 50 have very similar peak heights but it appears that Bayes IM is more variable than Bayes IM Comp 50.

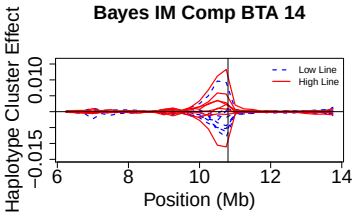
The A allele for this QTL is having a negative effect and the B allele has a positive effect. Additionally, the A allele frequency for the high line was 0.006

Figure 4.5.4: Haplotype effect estimates for BTA 14 between 6 and 14 MB



(a.)

(b.)



(c.)

and 0.995 for the low line. Again we except to see the haplotype clusters from the high line having a positive effect and the haplotype clusters from the low line having a negative effect. Figure 4.5.4 displays the haplotype cluster effect estimates for the three Bayes IM models.

For Bayes IM, there is only one haplotype cluster with a large positive effect. The remaining 15 haplotype clusters have very small to zero effect. Since we know that the A allele is having a negative effect, we should be seeing at least one haplotype cluster which has an obvious negative effect, which is not occurring. This is an indication that using 16 common haplotypes hinder our ability to detect the line specific QTL effects. Unlike the QTL on BTA 4, this QTL is being clearly detected by the Bayes IM model.

Bayes IM Comp is clearly detecting this QTL, as evident by the genetic variance plot above. At 10.8 MB on BTA 14, Bayes IM Comp has one high line and one low line haplotype clusters with a large positive effect. There are an additional three high line haplotype clusters with a small to medium positive effect. On the negative side, there is one high line cluster with a large negative effect and four low line haplotype clusters with small to medium sized negative effects. The remaining three high line and three low line haplotype clusters have no effect. As we mentioned before, the alleles for this QTL are not fixed within each line and it is therefore possible to have low line individuals with positive effect B alleles and high line individuals with negative effect A alleles. Thus the two haplotype clusters, which have an effect in the opposite direction from what was expected, are not concerning.

Finally, Bayes IM Comp 50 is clearly detecting this QTL, although all of the haplotype effects have been drastically reduced. There is one high line haplotype cluster with a large positive effect. An additional high line haplotype

cluster, low line haplotype cluster, and two common haplotype clusters have a small to medium positive effect. On the negative side, there are no haplotype clusters with a large effect, but there are two low line, one high line, and three common haplotype clusters with small negative effects. We would still expect to see haplotype clusters with larger negative effects. A possible explanation for this is that the inclusion of eight common haplotype clusters into the model may be too many for this data set and as a result the line specific effects are being muted.

4.5.3 Prediction Accuracy

We will now access the prediction accuracy of each model, as this is the best indicator for the best model. Since we are using simulated data, we are able to directly compare the estimated breeding value from our models to the true breeding value of the individual. Therefore, prediction accuracy is equal to $Corr\,(\mathbf{g}, \hat{\mathbf{g}})$, where $\mathbf{g}$ is the true breeding value and $\hat{\mathbf{g}}$ is the estimated breeding values. The standard error for this correlation is equal to:

$$\frac{1 - r^2}{\sqrt{n-2}},$$

where r is the correlation calculated above and n is equal to the number of individuals [31]. Each fold served as the evaluation set once and genetic correlations were predicted based on the pooled marker effects from using all folds except the evaluation fold. The correlations and standard errors for all three fold and all five models can be found in Table 4.5.3. The high fold, medium fold, and low fold refers to the folds made up of high, low, and medium phenotype individuals, respectively.

For all three folds the model with the highest prediction accuracy is Bayes

Table 4.5.3: Prediction Accuracy

	Evaluation Set		
Model	High Fold	Medium Fold	Low Fold
Bayes B	0.823 (0.005)	0.692 (0.007)	0.824 (0.005)
Bayes C	0.851 (0.005)	0.735 (0.006)	0.854 (0.005)
Bayes IM	0.836 (0.005)	0.716 (0.007)	0.847 (0.005)
Bayes IM Comp 50	0.791 (0.007)	0.697 (0.007)	0.802 (0.006)
Bayes IM Comp	0.603 (0.011)	0.630 (0.008)	0.602 (0.011)

a. Correlation (SE)

C, followed by Bayes IM, and then Bayes B. Additionally, the prediction accuracy was increased by including the eight common haplotype clusters into the Bayes IM Comp model. For all models except Bayes IM Comp, the ability to make predictions for the medium phenotype individuals was greatly decreased. For Bayes IM Comp, the medium fold is the fold that this model was best able to make predictions for. Bayes IM Comp is able to make better predictions when using the purer individuals to predict the individuals which are more composite in nature.

4.5.4 Conclusions

For this simulated data set, the traditional Bayes C model is outperforming the rest of the models. The Bayes C model had posterior values which were closest to how the data was generated. Using the genetic variance for 1 MB windows, Bayes C ranked the top five QTL all reasonably well and was able to precisely identify the true location of the QTLs. Finally, the prediction accuracy for Bayes C was the highest.

From a prediction standpoint, Bayes IM is outperforming the two Bayes IM Comp models. When it comes to QTL identification, Bayes IM Comp is performing better. In general, Bayes IM Comp ranks the top QTLs better than

Bayes IM. Bayes IM Comp also has much larger and easier to identify peaks for the haplotype effects and putative QTL effect estimates. The addition of the common haplotype cluster improves our prediction ability. However, our ability to identify the location of QTLs was reduced when the common haplotypes were included in the Bayes IM Comp model.

4.6 Overall Conclusions for Bayes IM Comp

From a prediction standpoint, there is little to any benefit to accounting for an individuals breed composition. We do just as well and in some cases better with a model using 100% common haplotypes as we did with a model using 50% common or 0% common haplotypes. Additionally, the haplotype models in general do not have better predictability than the Bayes B and Bayes C models, which are widely used.

The main advantage to the Bayes IM Comp model is in its ability to get clearer estimates of where the QTLs are occurring and within which line or breed the QTLs occurs. However, we must be very cautious in how we estimate our haplotype clusters. Inclusion of purebred individuals is important in getting accurate haplotype clusters for each breed. For the Simmental data set, we would be able to better control our haplotype clusters if we included purebred Hereford individuals and more purebred Angus individuals.

CHAPTER 5

EVALUATION OF BAYES IM PARENTAL COMP

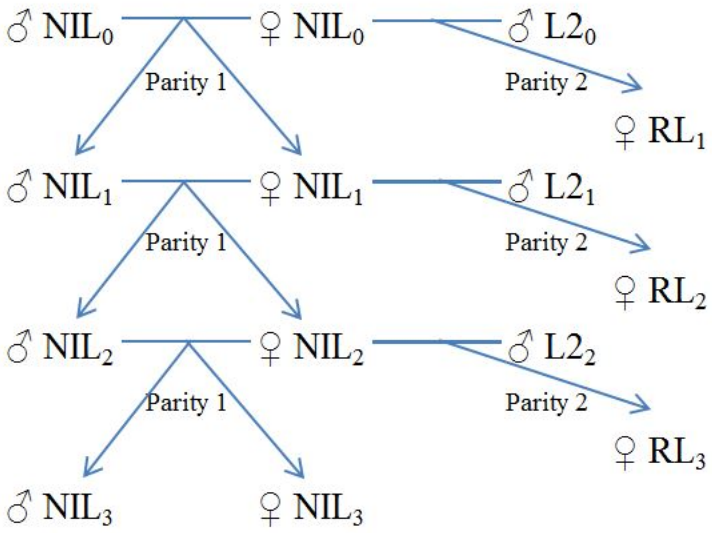
In experimental and production settings, two- and three-way crosses are quite common and in these situations the breed composition of the parents are quite different. A two-way cross is a cross between two lines and a three-way cross is a cross between the offspring from a two-way cross and an individual from a third line. Bayes IM Parental Comp was motivated by the two-way cross. However, Bayes IM Parental Comp would also be appropriate for populations where the parental breed composition is known and the paternal and maternal breed compositions are different from each other. We will be comparing Bayes IM Parental Comp to the traditional Bayes B and C models and to the basic Bayes IM model using a two-way cross population that the model was originally designed for. We will then be comparing Bayes IM Parental Comp to Bayes IM Comp using the simulated population introduced in Chapter 4, which will allow us to evaluate the impact of accounting for parental breed composition on prediction accuracy and QTL detection in a population that is not a two-way cross.

5.1 Reproductive Longevity Swine Data Set

5.1.1 Data Description

Data on 1,089 gilts from batches 5 through 14 of the Reproductive Longevity line (RL) at the University of Nebraska-Lincoln were used in this study. A batch is essentially a generation, except that a few generations where skipped due to space availability in the facilities used. The first four batches had sires from a different line and were excluded. The gilts were the progeny of a two-way cross between a dam on her second parity from the Nebraska Index Line (NIL) and a sire from the Landrace 2 line (L2). NIL individuals are from the University of Nebraska-Lincoln sow reproductive longevity resource population and are a composite population made up of "Landrace and Large White genetics which has been subjected to long-term selection for litter size since 1981" [63]. L2 individuals are a commercial Landrace sire line. The

Figure 5.1.1: Population Structure for the Reproductive Longevity Gilts



a. RL = Reproductive Longevity Line, NIL = Nebraska Index Line, L2 = Landrace 2 Line

breeding design for the reproductive longevity line gilts and the NIL line is illustrated in Figure 5.1.1.

Age of puberty was recorded for each gilt and was the trait of interest. Gilts which reach puberty at a young age tend to have a longer reproductive life as they tend to stay in the herd longer, produce more litters, and as a result give birth to more piglets [44, 53]. In addition, each individual was placed in a pen and diets were assigned to each pen. The diets include a standard diet of corn-soybean meal, an energy restricted diet of corn-soybean meal with a 20% caloric reduction, and an energy and amino acid restricted diet of corn-soybean meal with a 20% caloric reduction and a Lysine reduction [35]. The randomization of individuals to pens and diets to pens was approximately assigned as follows. First, individuals were randomly assigned to pens with the restriction that litter mates were not assigned to the same pen. Then diets were randomly assigned to pens with the restriction that no two adjacent pens were placed on the same diet.

The gilts were genotyped with the PorcineSNP60 BeadChip by Illumina [26]. In addition, SNP genotypes on L2 boars (n=24), NIL boars (n=22), and NIL sows (n=47) were used. The NIL sows are dams for the reproductive longevity gilts and the L2 boars are sires of batches 9 through 13 of the reproductive longevity population.

5.1.2 Models Compared

We compared a total of five models: Bayes B, Bayes C, Bayes IM (100% common model, this is the Bayes IM model with a total of 16 haplotype clusters shared by all lines), Bayes IM PC50 (50% common model, this is the Bayes IM Parental Comp model with 4 haplotype clusters assigned to NIL,

4 assigned to L2, and 8 assigned to be common among all lines for a total of 16 haplotype clusters), and Bayes IM PC (0% common model, this is the Bayes IM Parental Comp model with 8 haplotype clusters assigned to NIL, 8 haplotype clusters to L2, and zero haplotype clusters assigned to be common among all lines for a total of 16 haplotype clusters). For the Bayes IM models, fixed effects included an overall mean and an effect for diet by batch. Random effects included a sire effect, a litter by batch effect, and a pen by batch effect. All three of these random effects were assumed to be independent and identically distributed with a distribution that is normal with a mean of 0 and variances of σ^2_S, σ^2_L, and σ^2_P representing sire, litter by batch, and pen by batch respectively.

Because the GenSel software v4.73 [13] for the Bayes B and Bayes C models does not allow us to fit additional random effects into our model, Bayes B and C models considered all additional covariates as fixed effects. The models included fixed effects for for the overall mean, diet by batch, and pen by batch. Sire and litter by batch were not included in the models as this resulted in extremely decreased predictability. The other elements are defined as they were for the Bayes IM models.

5.1.3 Training and Evaluation Sets

All reproductive longevity individuals used in this study were comprised of 50% NIL and 50% L2 genetics. When estimating the haplotype clusters using these individuals, haplotype cluster labels for NIL and L2 are not identified. In addition to the genotypes of the 597 individuals from batches 5 through 10 of the RL line, genotypes for the 93 individuals of the two parental lines were used to train the HMM. The genotypes from the parental lines were included

to make the haplotype cluster line labels identifiable.

The training set only included the RL individuals from batches 5 through 10. This training set was used by all five models in order to estimate the marker effects. Prediction accuracy was evaluated on batches 11, 12, 13, and 14 separately to determine how our prediction accuracies decay as our prediction set gets further separated from our training set. Additionally, the training set was used to identify potential QTL with associations to age of puberty.

5.2 Results and Discussions for the NIL Swine Data Set

Evaluation of the reasonableness of the priors and the convergence of the Markov chain was done by examining the posterior distributions of the variance components and the trace plots. QTL identification was done by comparing the top 1 MB windows. We will also compare the estimated genetic variance at each SNP locus in the Bayes B and Bayes C models and each putative QTL locus in the Bayes IM versions. The haplotype cluster effect estimates for the three Bayes IM models will be compared. Prediction accuracy will be estimated and compared for each model.

5.2.1 Priors and Evaluation of Sample Convergence

Priors were based on previous studies preformed. To allow a fair comparison across models, the priors were chosen so that the prior values were consistent with the posterior means. Table 5.2.1 summarizes the prior, posterior mean, and standard error values for all the variance components used by each model. The prior for the genetic variance is only used as a prior for Bayes B and Bayes C models. The prior for haplotype effect variance, sire variance, pen by batch variance, and litter by batch variance were only used for the Bayes

IM models. Heritability for age of puberty is moderate in pigs with a typical value between 0.38 and 0.46 [33] and, therefore, we set a goal for 0.40. We used different residual variance priors for Bayes B and C and the Bayes IM models since their model equations differed. In general, all three Bayes IM models result in very similar posterior means for all parameters and the Bayes B and Bayes C models agree with each other.

For the residual variance and haplotype effect variances, the posterior means for each of the Bayes IM models are near the prior value. Bayes C and Bayes B estimate the residual variance to be a lot higher than any of the Bayes IM models, but the posterior means are close to the prior used for these models. The larger estimated residual variance in Bayes B and Bayes C is consistent with not including sire and pen by batch in the model. For genetic variance, Bayes B and Bayes C have posterior means very close to the prior of 100. In addition, the genetic variance is much lower than that estimated by the Bayes IM models and is consistent with what we have observed in both the Simmental and simulated data set. The higher residual variance and lower genetic variance seen within the Bayes B and Bayes C models leads to the lower heritability, which is below the range of 0.38 to 0.46, from previous studies. The Bayes IM models have heritabilities which fall within the range

Table 5.2.1: Prior and Posterior Means (SE) for Variance Components

Parameter	Priors	Posterior Mean				
		Bayes B	Bayes C	Bayes IM	Bayes IM PC 50	Bayes IM PC
Residual Variance	230/178	225.57 (18.81)	223.66 (24.34)	186.20 (27.41)	169.89 (28.14)	182.24 (26.94)
Haplotype Effect Variance	—/0.58	N/A	N/A	0.54 (0.16)	0.64 (0.18)	0.57 (0.16)
Genetic Variance	100/—	97.82 (15.627)	101.81 (29.92)	132.72 (33.83)	152.17 (35.40)	138.49 (33.09)
Heritability	0.4	0.30 (0.04)	0.31 (0.08)	0.393 (0.09)	0.45 (0.09)	0.41 (0.09)
Sire Variance	—/8	N/A	N/A	5.78 (5.12)	5.62 (4.60)	5.94 (4.98)
Pen by Batch Variance	—/8	N/A	N/A	6.97 (5.30)	6.94 (5.29)	6.61 (5.07)
Litter By Batch Variance	—/12	N/A	N/A	10.88 (7.86)	9.98 (7.44)	10.91 (7.91)

a. The first prior listed is for Bayes B and C, the second prior is for the Bayes IM models.

of 0.38 to 0.46. Bayes IM and Bayes IM PC agree, but Bayes IM PC 50 is estimating a lower residual variance and a higher genetic variance which leads to a heritability which is higher than any other model.

Based on the posterior distribution plots in Figure C.1.1 and Figure C.1.2 of Appendix C, the residual variance, haplotype effect variance, genetic variance, and heritability all appear to have mostly symmetric bell shaped distributions which are centered close to our prior value. The priors for sire variance, pen by batch variance, and litter by batch variances appear to be high. However, based on the distribution plots in Figure C.1.3, sire, pen, and litter have right skewed distributions with a large variance and our priors are contained nicely within their respective distributions. In addition to the posterior distribution plots, the trace plots were examined for all variance components. There were no trends or patterns observed within the trace plots. This indicated that the initial burn-in and overall number of MCMC samples are sufficient. Therefore the prior estimates appear to be reasonable and no further adjustments were made to the model.

5.2.2 QTL Identification and Haplotype effect estimates

The top ten 1 MB windows based on window genetic variance for all five models are summarized in Tables C.2.1 and C.2.2. Three regions showed up consistently across the five models. On chromosome 2 (SSC 2), the region between 12 and 15 MB showed up as the top QTL for Bayes IM PC. Bayes IM PC 50 ranked this QTL third and Bayes C ranked this QTL sixth. While Bayes B and Bayes IM did not rank this QTL in the top 10, Bayes B ranked this QTL 11th and Bayes IM ranked this QTL 16th, which is still a high ranked QTL. On SSC 6, the region between 87 and 89 MB showed up as a top 10

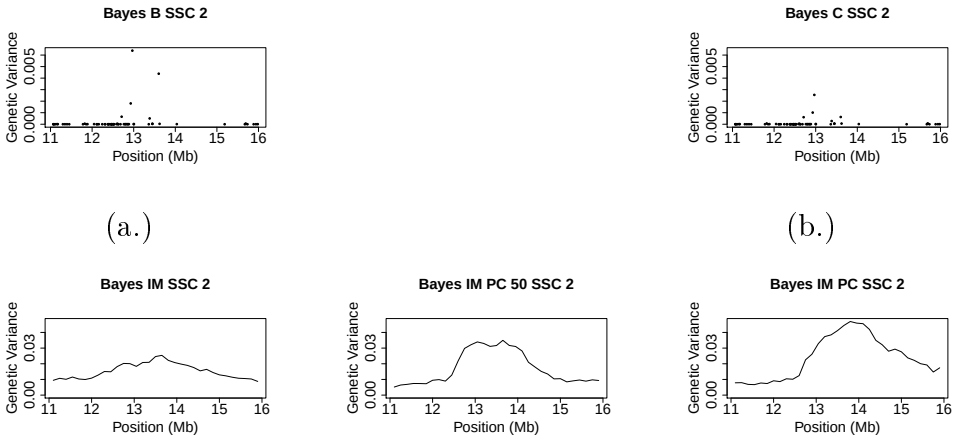
QTL for all models, except Bayes IM PC 50 which ranked this QTL 15th. On SSC 7, the region between 118 and 120 MB showed up as the top QTL for Bayes IM, Bayes IM PC 50, and Bayes B. Bayes C model ranked this QTL second and Bayes IM PC ranked this QTL 12th.

5.2.2.1 QTL on SSC 2

In this same population, Trenhaile et al. [63] discovered a QTL on SSC 2 between 13 and 14 MB. This region shows promise of having a functional QTL because it includes the candidate gene *P2X3R*, "which plays a role in implantation and sustained release of hormones associated with reproductive processes" [63] which could have an effect on age of puberty.

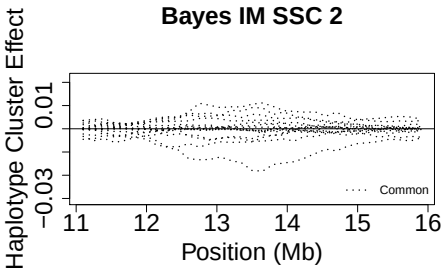
Figure 5.2.1 shows the genetic variances at each locus for SSC 2 between 11 and 16 MB for the five models examined. Bayes IM PC appears to do a better job detecting this QTL than either of the common haplotype models, Bayes IM or Bayes IM PC50, with Bayes IM PC50 having a slightly larger peak than Bayes IM. The Bayes B and Bayes C models detect this QTL well.

Figure 5.2.1: Genetic Variance for SSC 2 between 11 and 16 MB

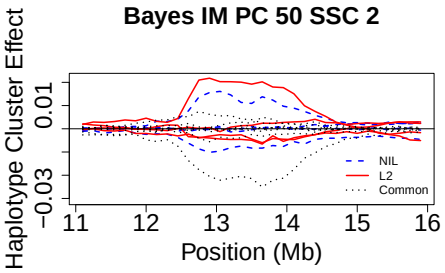


Even though this QTL is in the top ten for Bayes C but not for Bayes B, Bayes B is showing a slightly larger peak than Bayes C. This is consistent with trends observed within both the Simmental and simulated data set and it due to the thicker tails of the t-distribution used by the Bayes B model.

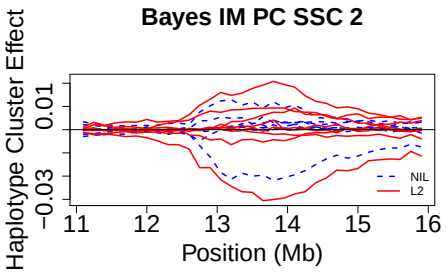
Figure 5.2.2: Haplotype Effect Estimates for SSC 2 between 11 and 16 MB



(a.)



(b.)



(c.)

The haplotype cluster effects for the three Bayes IM models can be seen in Figure 5.2.2. Bayes IM PC is indicating that there is a haplotype cluster assigned to NIL and one assigned to L2 which have a very large negative effect. Additionally, there is a haplotype cluster assigned to L2 which has a large positive effect. For Bayes IM PC 50, the large effect of the L2 cluster appears to be preserved and there is an additional NIL cluster with a large positive effect. The negative NIL and L2 clusters from Bayes IM PC seem to

be replaced by a common haplotype cluster within Bayes IM PC 50. Bayes IM is showing one negative cluster but there is no haplotype cluster which stands out on the positive side.

The probability of being in a given cluster within Bayes IM PC given you are in common haplotype cluster 14 within Bayes IM PC 50 was calculated, where common haplotype cluster 14 represents the large negative cluster observed in Figure 5.2.2 (b.). This revealed that there was a 21% probability of being in NIL haplotype cluster 3, a 7% probability of being in NIL haplotype cluster 6, a 9% probability of being in NIL haplotype cluster 8, 10% probability of being in L2 haplotype cluster 14, and a 32% probability of being in L2 haplotype cluster 16. L2 haplotype cluster 16 and NIL haplotype cluster 6 are the large negative clusters observed within Bayes IM PC. The effect size of common haplotype cluster 14 within Bayes IM PC 50 is smaller than the effect size of L2 cluster 16 within Bayes IM PC because the large negative effect of Bayes IM PC cluster 16 is being muted by the smaller effects from clusters 3, 9, and 14 within Bayes IM PC.

The probability of being in a given cluster within Bayes IM PC 50 given you are in L2 haplotype cluster 10 within Bayes IM PC was calculated, where L2 haplotype cluster 10 represents the large postie cluster observed in Figure 5.2.2 (c.). The probability was 27% of being in L2 cluster 6, 13% of being in L2 cluster 7, 19% of being in L2 cluster 8, 9% of being in common cluster 14, and 27% of being in common cluster 16. L2 cluster 6 represents the large positive L2 cluster and common cluster 14 represents the large negative cluster from the Bayes IM PC 50 model. The effect size of L2 haplotype cluster 10 within Bayes IM PC is smaller than the effect size of L2 haplotype cluster 6 within Bayes IM PC 50 because the large postie effect of Bayes IM PC 50 cluster 6

is being muted by the large negative effects from common cluster 16.

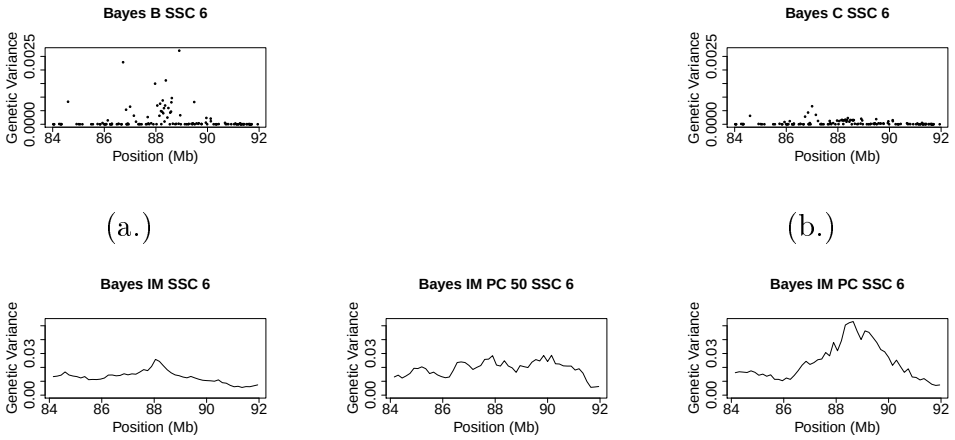
To measure the frequency of cluster membership, the expected frequency of cluster membership at a locus was calculated and is summarized in Table C.2.3. This revealed that five haplotype clusters from Bayes IM have expected frequencies of less than 0.01. In addition, there are two NIL haplotype cluster, one L2 haplotype cluster, and 4 common haplotype clusters from Bayes IM PC 50 and 3 NIL haplotype clusters and three L2 haplotype clusters from Bayes IM PC that have expected frequencies less than 0.01. This is an indication that at this particular location not all 16 haplotype clusters are needed. However, we cannot conclude that overall 16 haplotype clusters is too many since different haplotype clusters will be used in different locations.

5.2.2.2 QTL on SSC 6

The second location that we will look at is on SSC 6 between 87 and 89 MB. There are currently no studies which identify this exact location as a known QTL for age of puberty; however, there was one study which identified two different regions on SSC 6 as having a QTL for age of puberty. Nonneman et al [42] identified a QTL at 69 and 127 MB. Additionally, there are currently no known genes which are associated with age of puberty in the region between 87 and 89 MB.

Figure 5.2.3 shows the genetic variances at each locus for SSC 6 between 84 and 92 MB for the five models examined. Bayes IM PC is once again showing the largest spike for the potential QTL. Bayes IM shows a more defined peak at 88 MB, whereas Bayes IM PC 50 seems to be unsure of the true location of the QTL since it is showing an elevated genetic variance for values between 86 and 92 MB. Bayes B is detecting a much larger peak and a larger number of

Figure 5.2.3: Genetic Variance for SSC 6 between 84 and 92 MB

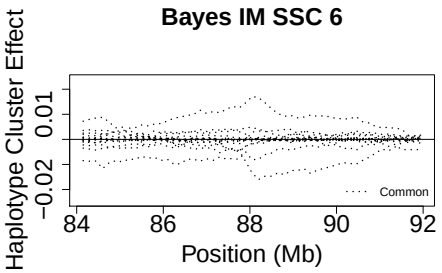


SNPs with an effect than the Bayes C model. The larger peak can be explained by the thicker tails of the t-distribution, but the larger number of SNPs overall is in indication that Bayes B is better detecting this QTL than Bayes C.

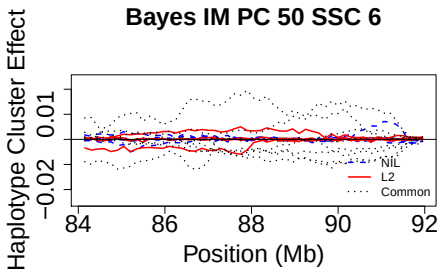
The haplotype cluster effects for the three Bayes IM models are presented in Figure 5.2.4. Bayes IM PC is showing that at 88.5 MB there is one NIL and one L2 haplotype cluster with a large positive effect and one NIL and one L2 haplotypes cluster with a large negative effect. Bayes IM PC50 is showing one common haplotype cluster with a large positive effect between 86.25 and 88.5 MB and there are no haplotype clusters which stand out on the positive side. Bayes IM is showing one common haplotype cluster with a large effect at 88 MB and a common haplotype cluster with a large negative effect starting at 88.5 MB.

The large negative L2 and large negative NIL haplotype cluster within the Bayes IM PC model are not present in the other two models. In each case, this is because the large effects of these clusters are being spread among multiple clusters in the other models. For example, we calculated the probability of an

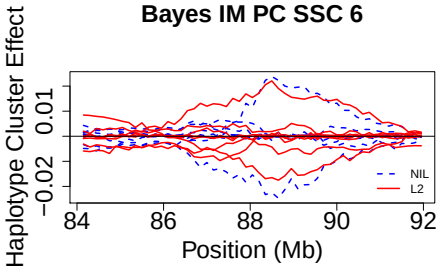
Figure 5.2.4: Haplotype Effect Estimates for SSC 6 between 84 and 92 MB



(a.)



(b.)



(c.)

individual being in a particular cluster within Bayes IM given there were in NIL cluster 1 within Bayes IM PC, which represents the NIL cluster with a large negative effect. This revealed that the probability of being in cluster 3 was 23%, the probability of being in cluster 12 was 37%, and the probability of being in cluster 13 was 23%. Thus the large negative effect is being spread among three different clusters within Bayes IM. Additionally, the negative effect in Bayes IM is from cluster 12, which is the cluster with the strongest map to Bayes IM PC cluster 1, which explains the much smaller effect observed within Bayes IM.

To measure the frequency of cluster membership, the expected frequency of cluster membership at a locus was calculated and is summarized in Table

C.2.4. Table C.2.4 revealed that eight haplotype clusters from Bayes IM have expected frequencies of less than 0.01. There are three NIL, two L2, and two common haplotype clusters in Bayes IM PC 50 and five NIL and three L2 haplotype clusters in Bayes IM PC that have expected frequencies less than 0.01. Again, at this location 16 haplotype clusters are not needed and there are more unused haplotype clusters at SSC 6 than at SSC 2.

5.2.2.3 QTL on SSC 7

Currently there are zero studies which identify a QTL on SSC 7 between 118 and 120 Mb for age of puberty. However, there are several studies which identify different regions on chromosome 7 as having a QTL for age of puberty. Cassady et al. [8] identified two QTL on chromosome 7, one at 1 MB and one at 58 MB. Yang et al. [66] identified a QTL at 54 MB and Nonneman et al [42] identified a QTL at 43 and 75 MB. Similar to the QTL on SSC 6, there are currently no known genes which are associated with age of puberty on SSC 7 between 118 and 120 MB.

Figure 5.2.5: Genetic Variance for SSC 7 between 115 and 122 MB

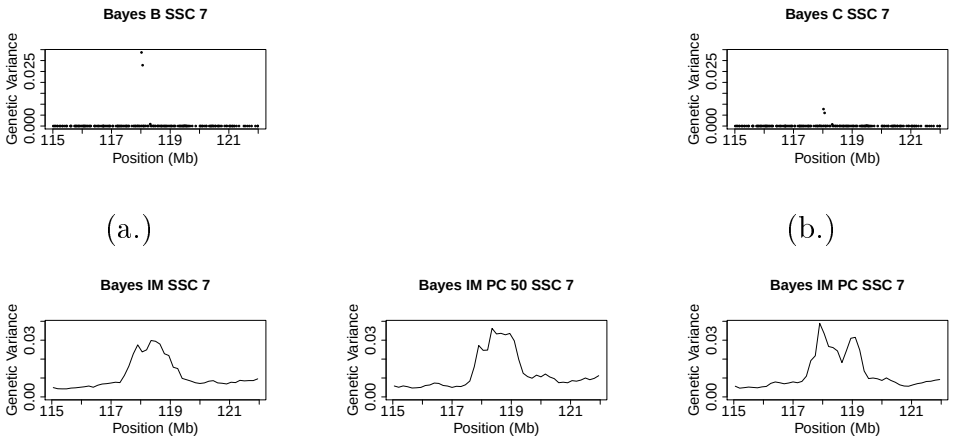
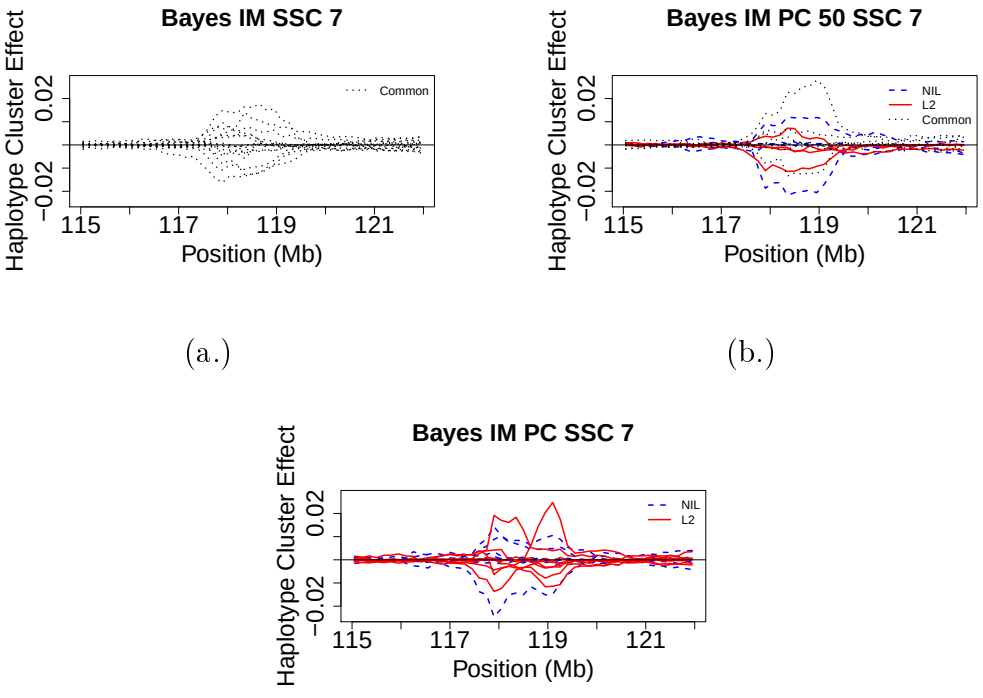


Figure 5.2.5 shows the genetic variances at each locus for the five models examined. Recall, that Bayes IM and Bayes B identified this region as their top QTL and Bayes C identified this as the second ranked QTL. Even though Bayes IM ranked this QTL higher than either Bayes IM PC model, both Bayes IM PC models have a larger peak than Bayes IM. In addition, Bayes IM PC is showing two peaks rather than a single peak which is likely why this QTL was not in the top 10 for that model. For this QTL, Bayes B and Bayes C show obvious elevations in the genetic variance around 118 MB. With Bayes B having a much larger elevation than Bayes C, which is due to the thicker tailed t-distribution used by Bayes B.

The haplotype cluster effect estimates for the three Bayes IM models are

Figure 5.2.6: Haplotype Effect Estimates for SSC 7 between 115 and 122 MB



(a.)

(b.)

(c.)

presented in Figure 5.2.6. Bayes IM PC is showing one L2 haplotype cluster with a large positive effect around 118 MB and a different L2 haplotype cluster with a large positive effect around 119 MB, which explains the two peaks we saw above in Figure 5.2.5. Additionally, there is one NIL haplotype cluster with a large negative effect which peaks around 118 MB. When we look at the Bayes IM PC 50 model, the two peaks have disappeared and we are seeing one clearly defined peak with one common haplotype cluster with a large positive effect and one NIL haplotype cluster with a large negative effect. Finally, Bayes IM shows elevations in cluster effects but there is not one haplotype cluster which stands out above the rest in either direction.

We have seen plenty of evidence that shows muted effects, like those seen in Bayes IM, occur when large effect clusters have effects which are spread among multiple clusters of another model. Here we want to provide evidence that two models have effects which agree when a majority of the effect from one cluster is being placed into a cluster from another model. For example, we calculated the probability of being in a particular cluster within Bayes IM PC given we were in NIL haplotype cluster 2 in Bayes IM PC 50, which represents the large negative effect cluster. The probability of being in Bayes IM PC NIL cluster 1 was 12%, the probability of being in NIL cluster 3 was 9%, and the probability of being in NIL cluster 8 was 70%, where NIL cluster 8 represents the NIL cluster with a large negative effect in Bayes IM PC. Thus, almost all of the negative effects from NIL cluster 2 in Bayes IM PC 50 are being placed into NIL cluster 8 within Bayes IM PC, which explain why these two clusters have very similar effects.

To measure the frequency of cluster membership, the expected frequency of cluster membership at a locus was calculated and is summarized in Table

C.2.5. Table C.2.5 reveals that three haplotype clusters from Bayes IM, two NIL haplotype clusters and three common haplotype clusters from Bayes IM PC 50, and three NIL haplotype clusters and two L2 haplotype clusters from Bayes IM PC have expected cluster membership frequencies which are less than 0.01. At SSC 7, there are fewer haplotype clusters with low membership probabilities than there was at SSC 2 and SSC 6, but there is still in indication at this location that 16 haplotype clusters is too many.

We observed two trends within the QTL investigation. In general, the Bayes IM PC models result in clearer and more pronounced QTL effect estimates than Bayes IM. This is consistent with what was observed in both the Simmental and the simulated data set. Additionally, the Bayes IM PC models are able to better detect the line specific haplotype effect estimates since the Bayes IM haplotype effects are being muted when multiple clusters are combined together, which is also consistent with the smaller QTL effects observed in the genetic variance plots for Bayes IM.

5.2.3 Prediction Accuracy

The prediction accuracy of the five models were compared. Using the three Bayes IM models above, we estimated each individual i's breeding value, $\hat{g}_i = \sum_j M_{ij}\hat{b_j}$, where M_{ij} is the genotype covariate for individual i and putative QTL j and $\hat{b_j}$ is the estimated effect for putative QTL j from the model. Next, we considered the bi-variate distribution for y_i and $\hat{g}_i$. This was assumed to be normally distributed as follows:

$$\begin{pmatrix} y_i \\ \hat{g}_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 + d_i \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\right),$$

where y_i represented the age of puberty for individual i, d_i represents the effect of diet by replication, and μ_1 and μ_2 are the overall means for y_i and $\hat{g}_i$ respectively. Prediction accuracy was measured as the correlation between the residuals and is equal to:

$$\frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}},$$

where the 1 subscript represents y_i and the 2 subscript represents $\hat{g}_i$. The only change for the Bayes B and Bayes C models is what is represented by j. For Bayes B and Bayes C, j represents the j^{th} SNP rather than the j^{th} putative QTL. The correlations and standard errors are given below in Table 5.2.2 and were estimated using SAS/STAT software, Version 9.4 of the SAS System for Windows [55].

Table 5.2.2: Prediction Accuracies for Age of Puberty

Batch	11	12	13	14
Bayes B	0.302 (0.081)	0.198 (0.087)	0.193 (0.097)	-0.042 (0.103)
Bayes C	0.317 (0.080)	0.231 (0.086)	0.177 (0.097)	-0.024 (0.103)
Bayes IM	0.346 (0.078)	0.232 (0.086)	0.084 (0.100)	0.097 (0.102)
Bayes IM PC 50	0.357 (0.078)	0.163 (0.088)	0.113 (0.099)	0.063 (0.103)
Bayes IM PC	0.317 (0.080)	0.157 (0.089)	0.110 (0.099)	0.041 (0.103)

a. Residual Correlation (SE)

As expected, the prediction accuracies for all the models decreased the farther the evaluation population was from the training population. Bayes IM PC 50 had the highest prediction accuracy for batch 11. Bayes IM and Bayes C have the highest prediction accuracies for batch 12. Bayes B has a slightly higher prediction accuracy with batch 13. By batch 14, all the models had a prediction accuracy which was close to zero.

Kachman [29] has observed that Bayes B and Bayes C have higher prediction accuracies compared to Bayes IM in a homogeneous population and,

therefore, it is promising to see that all three Bayes IM models have similar prediction accuracies and in some cases higher prediction accuracies than Bayes B and Bayes C for batch 11. Comparing just the Bayes IM models, we can see that the Bayes IM PC 50 model has a higher prediction accuracy than the Bayes IM PC model for every batch. Additionally, for batches 11 and 13, Bayes IM PC 50 has a higher prediction accuracy than the Bayes IM model. Bayes IM has a much larger drop in prediction accuracy that either Bayes IM PC or Bayes IM PC 50 for batch 13 and the Bayes IM PC models seem to maintain predictability longer than Bayes IM does.

5.2.4 Conclusions

The number of rare haplotype clusters suggests that a haplotype model with a fewer number of overall haplotype clusters should be considered. Since the NIL line is made up of Landrace and large white genetics, we should consider a model with fewer NIL specific haplotype clusters and fewer common haplotype clusters, since these were the clusters that tended to have lower expected cluster membership frequencies in the three cases that we examined. When it came to QTL detection, all five models were able to detect the presence of each QTL we examined. However, the Bayes IM PC models were able to do a better job detecting the presence of these QTL than the Bayes IM model. The inclusion of the common haplotype clusters to the Bayes IM PC model appeared to help stabilize the QTL effect estimates in the case of the QTL on SSC 7. The common haplotype clusters appeared to hurt our QTL detection for SSC 6, but this may not be the case if we use a model with fewer common haplotype clusters.

As for prediction accuracy, the inclusion of the common haplotypes to the

Bayes IM PC model improved our prediction accuracy. Without the common haplotype clusters, Bayes IM PC is preforming below Bayes IM. Overall, the best preforming model appears to be the Bayes IM PC 50 model.

5.3 Evaluation of Bayes IM Parental Comp Using A Simulated Data Set

In chapter 4, we investigated the individual breed composition version on a simulated data set. The simulated data set from chapter 4 will be used to evaluate the parental breed composition model relative to the individual breed composition model. We will be reporting the posterior means, QTL rankings, haplotype cluster plots, and prediction accuracies for four models. The Bayes IM Comp models were presented previously in Chapter 4 and are being shown here in order to make the comparisons easier. The second two models are: Bayes IM PC (0% common, this is the Bayes IM Parental Comp model with 8 haplotype clusters assigned to the low line, 8 haplotype clusters to the high line, and zero haplotype clusters assigned to be common among both lines for a total of 16 haplotype clusters) and Bayes IM PC 50 (50% common, this is the Bayes IM Parental Comp model with 4 haplotype clusters assigned to the low line, 4 assigned to the high line, and 8 assigned to be common among both lines for a total of 16 haplotype clusters). The Bayes IM PC models use the same model equation and training and evaluations sets as Bayes IM Comp, which were reported in Chapter 4.

5.3.1 Posterior Distribution Comparison

Recall from chapter 4 that Bayes IM Comp had posterior means that were much more extreme than any either Bayes B, Bayes C, or Bayes IM. The

Table 5.3.1: Prior and Posterior Means (SE) for Variance Components

Model	Genetic Variance	Residual Variance	Heritability	Haplotype Effect Variance
Prior	0.4	0.6	0.4	0.0003
Bayes IM Comp[a]	0.539 (0.019)	0.438 (0.019)	0.552 (0.019)	0.00040 (0.00004)
Bayes IM Comp 50[a]	0.496 (0.013)	0.480 (0.011)	0.508 (0.011)	0.00033 (0.00003)
Bayes IM PC	0.579 (0.025)	0.398 (0.025)	0.592 (0.025)	0.00048 (0.00005)
Bayes IM PC 50	0.567 (0.020)	0.409 (0.019)	0.581 (0.019)	0.00049 (0.00005)

a. Results were previously reported in Table 4.5.1.

posterior means for the Bayes IM PC models are even more extreme than those estimated by Bayes IM Comp. The posterior distribution plots for Bayes IM PC (Figure B.2.3), does not have a nice symmetric bells shaped curve for the genetic variance, residual variance, and heritability. Instead there is a bi-modal curve. This is an indication that either a larger burn-in or a larger number of MCMC samples may be needed. The posterior distribution plots for Bayes IM PC 50 (Figure B.2.3) does not show a bi-model distribution; however, the posterior distributions for this model as slightly skewed and a larger burn-in or a larger number of MCMC samples may help this model as well.

To help confirm the need for a larger number of MCMC samples, the trace plots were examined. For Bayes IM PC and Bayes IM PC 50 the genetic variance and heritability trace plots are showing an upward trend. Similar to what was observed for Bayes IM Comp and Bayes IM Comp 50, the residual variance trace plots for the two Bayes IM PC models are showing a downward trend. This is a good indication that the overall number of samples is not large enough for the model to settle.

5.3.2 QTL Identification and Haplotype Effect Estimates

To determine if there was an improvement in QTL identification by including the parental breed composition instead of the individual breed composition,

the top five QTLs and the ranks in terms of genetic variance for the two Bayes IM Comp and Bayes IM PC models are presented below in Table 5.3.2. Bayes IM PC is doing worse than Bayes IM Comp, but there is a slight improvement when the common haplotypes are included in the model since Bayes IM PC 50 is estimating the fourth and fifth ranked QTLs better than Bayes IM Comp 50 did. However, overall Bayes IM Comp ranks the top QTLs closer to the truth and it appears there is little benefit to using the parental breed composition when it comes to overall QTL identification.

Table 5.3.2: Top QTLs for the Simulated Data Set

			A Allele Freq.		Rank			
BTA	Pos.	GenVar	High Line	Low Line	Bayes IM Comp	Bayes IM Comp 50	Bayes IM PC	Bayes IM PC 50
4	6.6	0.00086	0.002	0.998	25	51	102	51*
28	10.8	0.00084	0.994	0.003	16**	99	126	141**
5	114	0.00069	0.996	0.002	4	2*	4*	2*
14	10.8	0.00068	0.006	0.995	2	13	3	3*
1	113	0.00054	0.005	0.980	5*	10*	7*	5*

a. * MB below, **MB above

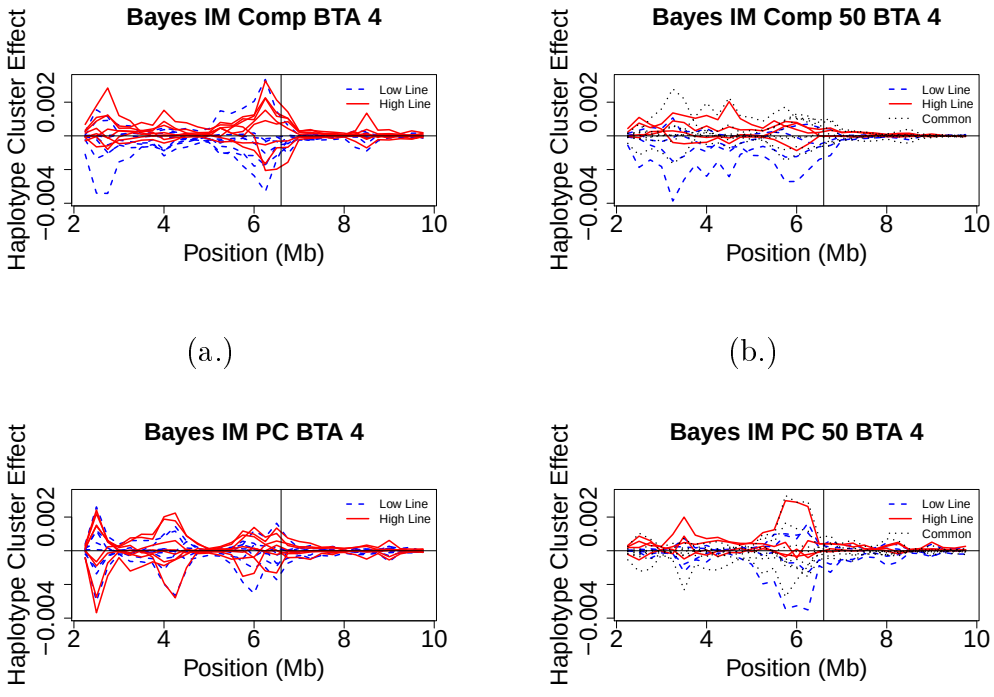
As in chapter 4, we are going to examine the top QTL on BTA 4 and the fourth ranked QTL on BTA 14. We will be focusing specifically on the haplotype cluster estimates as these are of the greatest interest. However, the genetic variances at each locus for the two Bayes IM PC models can be seen in Figures B.3.1 and B.3.2. We hope to see if the parental version had fewer haplotype clusters which were estimated to have an effect in the opposite direction from what was expected.

5.3.2.1 QTL on BTA 4

Recall, for the QTL on BTA 4, the A allele is having a negative effect, while the B allele has a positive effect. Thus we would expect to see the haplotype clusters from the high line having a positive effect and the haplotype clusters from the low line having a negative effect, since the A allele frequency for

the high line was 0.002 and the A allele frequency for the low line was 0.998. Figure 5.3.1 shows the haplotype cluster effects for the two Bayes IM Comp models and the two Bayes IM PC models.

Figure 5.3.1: Haplotype Effect Estimates for BTA 4 Between 2 and 10 MB



(a.) (b.)

(c.) (d.)

Similar to what we observed with Bayes IM Comp, Bayes IM PC is not clearly detecting this QTL. There are positive effect estimates showing possible QTLs nears 2 and 4 MB even though there is no true QTL at either location. As for the haplotype effect estimates themselves we are going to focus our discussion on what is being observed at 6 MB. Bayes IM PC has one low line and three high line haplotype clusters with a positive effect and one high line and three low line haplotype cluster with a negative effect. The remaining eight haplotype clusters have effects which are close to 0. This is consistent

with the Bayes IM Comp haplotype cluster effects. We still observe a negative high lines and positive low lines. There are a few high line individuals with the A allele rather than the B allele and a few low line individuals with the B allele rather than the A allele which results in haplotype clusters having the opposite effect from what we would expect. As evidence of this, low line cluster 7 in Bayes IM Comp and low line cluster 2 in Bayes IM PC represent the low line clusters with a large positive effect. The expected cluster membership probability is 0.0024 for Bayes IM Comp low line cluster 7 and 0.0004 for Bayes IM PC low line cluster 2, which confirms that these haplotype cluster represent a few low line individuals with B alleles.

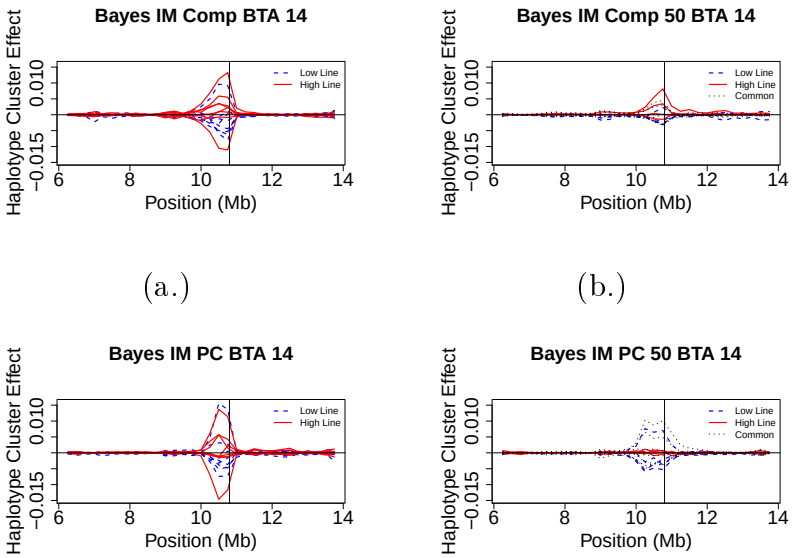
Of the four models, Bayes IM PC 50 is showing the clearest QTL around 6 MB. There is still a slight elevation around 4 but this elevation is smaller than what was observed for the other three models. Bayes IM PC 50 is showing one common and one high line haplotype cluster with a large positive effect and one common and two low line haplotype clusters with a medium positive effect. Additionally, there is one low line and one common haplotype cluster with a large negative effect and three common and one low line haplotype cluster with a medium negative effect.

5.3.2.2 QTL on BTA 14

Bayes IM Comp, Bayes IM PC, and Bayes IM PC 50 all estimate the QTL on BTA 14 equally well. Bayes IM Comp 50 is not estimating this QTL quite as well but it is still ranked within the top 15. Similar to the QTL on BTA 4, we expect to observe the high line haplotype clusters having a positive effect and the low line haplotype clusters having a negative effect. This is because the A allele for this QTL is having a negative effect with a frequency of 0.995 within

the low line and the B allele has a positive effect with a frequency of 0.994 in the high line. Figure 5.3.2 displays the haplotype cluster effect estimates for the two Bayes IM Comp models from chapter 4 and the two Bayes IM PC models..

Figure 5.3.2: Haplotype effect estimates for BTA 14 between 6 and 14 MB



The haplotype cluster effect estimates are very similar between Bayes IM PC and Bayes IM Comp. Both models have one low line haplotype cluster which is showing a large positive effect and one high line cluster which is showing a large negative effect. These clusters do not agree with what we expect, however, since this is a trend across two separate models it is not an issue with the model. Rather this is the result of a few individuals in the high line with A alleles and a few individuals in the low line with B alleles. In addition, Bayes IM PC is showing one high line haplotype cluster with a large positive effect and two high line haplotype clusters and one low line haplotype cluster with a small to medium positive effect. On the negative side there are

five low line haplotype clusters with small to medium effects.

The Bayes IM PC 50 is showing two common and one low line haplotype clusters with a medium positive effect and three common and three low line haplotype clusters with small to medium negative effects. This is slightly different from Bayes IM Comp 50 which showed no positive low line clusters and had a large positive high line cluster. These differences are a result of how the common haplotype clusters were formed within the two models.

5.3.3 Prediction Accuracy

Next we will evaluate the impact on prediction accuracy of including the parental breed composition rather than an individual breed composition. As we did in Chapter 4, we measure prediction accuracy as the $Corr\,(\mathbf{g}, \hat{\mathbf{g}})$, where $\mathbf{g}$ is the true breeding value and $\hat{\mathbf{g}}$ is the estimated breeding values. Table 5.3.3 summarizes the prediction accuracy and standard errors for the Bayes IM PC models.

Table 5.3.3: Prediction Accuracy (SE)

	Evaluation Set		
Model	High Fold	Medium Fold	Low Fold
Bayes IM Comp [a]	0.603 (0.011)	0.630 (0.008)	0.602 (0.011)
Bayes IM Comp 50[a]	0.791 (0.007)	0.697 (0.007)	0.802 (0.006)
Bayes IM PC	0.761 (0.007)	0.628 (0.008)	0.759 (0.007)
Bayes IM PC 50	0.804 (0.006)	0.678 (0.007)	0.801 (0.006)

a. Results were previously reported in Table 4.5.2.

For the high fold and low fold, the use of the parental breed composition was an improvement over the use of the individuals breed composition. For the high fold, the predictors were trained using individuals from the low and medium folds. Individuals in the medium fold represent individuals whose maternal and paternal breed composition is different from each other. When

the individuals used in training have maternal and paternal breed compositions which differ from each other is is better to use the parental breed composition rather than the individual breed composition.

For the medium fold, the use of the parental breed composition was not an improvement. The individuals in the training set consisted of individuals from the high and low fold which had an average dominant breed composition of 85%. Thus, there is limited information gained by using the parental breed composition instead of the individual breed composition since the maternal and paternal breed composition of these individuals are likely to be very similar. When the breed composition of the individuals used in training does not differ from the breed composition of the parents it is simpler to use the individual breed composition.

For all three folds the inclusion of the common haplotype clusters into the Bayes IM PC model improved our predictability. However, the prediction accuracy is still not greater than Bayes IM, Bayes B, or Bayes C. For this data set in particular, the simpler SNP based model outperforms the more complicated models.

5.3.4 Conclusions

The posterior estimates differed from our expectation, but the data itself was simulated using a SNP based genetic variance which differs from a haplotype based genetic variance. Thus, our haplotype based models may be doing a better job of detecting the overall genetic variance. In the case of the model which included common haplotype clusters, there was an advantage when it came to QTL identification. Overall Bayes IM PC 50 ranked the QTLs better than Bayes IM Comp 50 and clearly identified the QTL on both BTA 4 and

BTA 14. Also, using the parental breed composition rather than an individual breed composition is helpful as long as not all the individuals used in training have maternal and paternal breed compositions which differ from each other.

As these populations are only separated from the ancestral population by 20 generations, it is possible that including a larger number of common haplotype clusters may better capture the true haplotype diversity. For example, a 75% common model would assign two haplotype clusters to the low line, two to the high line, and 12 to be common among both lines. Additionally, since including more common haplotypes improves the prediction, we may be able to get closer to the prediction accuracy of Bayes IM but still be able to glean some line specific information out of the haplotype cluster estimates.

5.4 Overall Conclusions for Bayes IM Parental Comp

Both data sets used to evaluate the Bayes IM Parental Comp models reveal one commonality. There is an advantage to accounting for haplotype clusters which were common among both lines. As noted before, for the reproductive longevity data set the NIL line is made up of large white and Landrace genetics and thus the inclusion of common haplotype clusters makes sense. Additionally, the two lines from the simulated data set were created from the same ancestral population and the populations were only separated by 20 generations. Thus, they are likely to contain many similar haplotypes. However, the inclusion of common haplotype clusters could depend on the underlying genetic architecture.

CHAPTER 6

CONCLUSIONS AND FUTURE RESEARCH

In general, SNP based models outperform the haplotype based models in this study. Bayes C, specifically, had the highest prediction accuracy for a majority of the data sets and cases investigated above. Bayes IM Comp and Bayes IM Parental Comp do have promise in being able to identify the breed specific haplotype clusters which have the largest effect on the phenotype. However, this advantage comes at the cost of the overall computing time that these models require.

We examined the computing time for the simulated data set as this was the only data set which was ran on every model considered above. For this data set there were 12,500 individuals and approximately 60,000 SNPs. Bayes B, Bayes C, Bayes IM, Bayes IM Comp, Bayes IM Comp 50, Bayes IM PC, and Bayes IM PC 50 were run for a total of 42,000 MCMC iterations. Table 6.0.1 summarizes the total computing time for the seven models considered.

Unlike Bayes B and Bayes C, all versions of Bayes IM are able to run in parallel which greatly decreases the overall time these models take to run. However, even running in parallel, the Bayes IM models all have greatly increased computing time when compared to Bayes B and Bayes C. Bayes IM takes approximately three times as long as Bayes C. Bayes IM Comp takes almost seven times as long as Bayes C. Bayes IM Parental Comp is the most

Table 6.0.1: Computing Time for Simulated Data Set

Model	Computing Time	CPUs Used
Bayes B	8:01:57	1
Bayes C	6:48:17	1
Bayes IM	18:54:55	16
Bayes IM Comp	40:29:58	16
Bayes IM Comp 50	41:48:03	16
Bayes IM PC	63:53:22	16
Bayes IM PC 50	60:49:01	16

a. Time in hh:mm:ss

elaborate of the models and it takes approximately ten times as long as Bayes C. There are certainly ways to speed up the algorithms and increase the efficiency of the Bayes IM models and, in order for the models to be utilized in an industry setting, this needs to be done. As computers improve over time, the increased computing time may become less of an issue.

Two additional extensions to the Bayes IM model should be considered. The first extension is to include information on putative functional variants in the model. When putative functional variants have been identified it is typically the case that the vast majority of individuals are not genotyped for the putative functional variant alleles. By using haplotype clusters it is hoped that the model will be better able to predict which putative functional variant alleles an individual has based on local haplotype information.

The second extension is to incorporate information on individuals who have not been genotyped using pedigree relationships. By also including pedigree relationships, we can directly incorporate phenotypic information on individuals who have not been genotyped. In addition, including pedigree information will allow us to predict the haplotype clusters for both individuals with and without a genotype.

BIBLIOGRAPHY

[1] I. Aguilar, I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Science*, 93(2):743–752, 2010.

[2] L. Andersson. Genetic dissection of phenotypic diversity in farm animals. *Nature Reviews Genetics*, 2(2):130–138, 2001.

[3] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37 (6):1554–1563, 1966.

[4] A. Blasco. The Bayesian controversy in animal breeding. *Journal of Animal Science*, 79(8):2023–2046, 2001.

[5] S. Bolormaa, J.E. Pryce, K. Kemper, K. Savin, B.J. Hayes, W. Barendse, Y. Zhang, C.M. Reich, B.A. Mason, R.J. Bunch, B.E. Harrison, A. Reverter, R.M. Herd, B. Tier, H.-U. Graser, and M.E. Goddard. Accuracy of prediction of genomic breeding values for residual feed intake and carcass and meat quality traits in Bos taurus, Bos indicus, and composite beef cattle. *Journal of Animal Science*, 91(7): 3088–3104, 2013.

[6] F.V. Brito, J.B. Neto, M. Sargolzaei, J.A. Cobuci, and F.S. Schenkel. Accuracy of genomic selection in simulated populations mimicking the extent of linkage disequilibrium in beef cattle. *BMC genetics*, 12(1):80, 2011.

[7] A. Carvajal-Rodríguez. Simulation of genomes: a review. *Current genomics*, 9(3):155–159, 2008.

[8] J.P. Cassady, R.K. Johnson, D. Pomp, G.A. Rohrer, L.D Van Vleck, E.K. Spiegel, and K.M. Gilson. Identification of quantitative trait loci affecting reproduction in pigs. *Journal of Animal Science*, 79(3): 623–633, 2001.

[9] S.A. Clark, J.M. Hickey, and J.H. van der Werf. Different models of genetic variation and their effect on genomic evaluation. *Genetics Selection Evolution*, 43(18), 2011. doi: 10.1186/1297-9686-43-18.

[10] H.D. Daetwyler, R. Pong-Wong, B. Villanueva, and J.A. Woolliams. The impact of genetic architecture on genome-wide evaluation methods. *Genetics*, 185(3):1021–1031, 2010.

[11] T. Druet and F. Farnir. Use of ancestral haplotypes in genome-wide association studies. In C. Gondro, J. Werf, and B. Hayes, editors, *Genome-Wide Association Studies and Genomic Prediction*, pages 347–380. Springer, 2013.

[12] R. Durbin. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[13] R.L. Fernando and D.J. Garrick. GenSel-user manual for a portfolio of genomic selection related analyses. *Animal Breeding and Genetics, Iowa State University, Ames*, 2008.

[14] R.L. Fernando and D.J. Garrick. Bayesian methods applied to GWAS. In C. Gondro, J. Werf, and B Hayes, editors, *Genome-Wide Association Studies and Genomic Prediction*, pages 237–274. Springer, 2013.

[15] R.L. Fernando, J.C. Dekkers, and D.J. Garrick. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole genome analyses. preprint (2014).

[16] D.J. Garrick, J.F. Taylor, and R.L. Fernando. Deregressing estimated breeding values and weighting information for genomic regression analyses. *Genetics Selection Evolution*, 41(1):1, 2009.

[17] D. Gianola and J.L. Foulley. Non-linear prediction of latent genetic liability with binary expression: an empirical Bayes approach. In *Proceedings of the 2nd World Congress on Genetics Applied to Livestock Production*, volume 7, pages 293–303, Madrid, Spain, 1982.

[18] A.R. Gilmour, B.J. Gogel, B.R. Cullis, S.J. Welham, R. Thompson, D. Butler, M. Cherry, D. Collins, G. Dutkowski, S.A. Harding, K. Haskard, A. Kelly, S.G. Nielsen, A. Smith, Verbyla A.P., and I.M.S. White. ASReml user guide. release 4.1 structural specification. *VSN International Ltd*, 2014.

[19] M.E. Goddard and B.J. Hayes. Genomic selection. *Journal of Animal breeding and Genetics*, 124(6):323–330, 2007.

[20] Y. Guan and M. Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815, 2011.

[21] D. Habier, R.L. Fernando, K. Kizilkaya, and D.J. Garrick. Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, 12 (186), 2011. doi: 10.1186/1471-2105-12-186.

[22] B.J. Hayes, P.J. Bowman, A.C. Chamberlain, K. Verbyla, and M.E. Goddard. Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution*, 41(1):51, 2009.

[23] B.J. Hayes, P.J. Bowman, H.D. Daetwyler, J.W. Kijas, and J.H. van der Werf. Accuracy of genotype imputation in sheep breeds. *Animal Genetics*, 43(1):72–80, 2011.

[24] C.R. Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447, 1975.

[25] Illumina Inc. BovineSNP50 v3 DNA Analysis BeadChip product description, 2016. URL `https://www.illumina.com/products/by-type/microarray-kits/bovine-snp50.html`.

[26] Illumina Inc. PorcineSNP60 v2 DNA Analysis BeadChip product description, 2017. URL `https://www.illumina.com/products/by-type/microarray-kits/porcine-snp60.html`.

[27] J. Johnston, G. Kistemaker, and P.G. Sullivan. Comparison of different imputation methods. In *Proceedings of the 2011 Interbull meeting*, volume 44, pages 25–33, Stavanger, Norway, 2011.

[28] S.D. Kachman. Genomic prediction model based on haplotype clusters. In *Joint Statistical Meetings*, Seattle, WA, August 2015.

[29] S.D. Kachman. Genomic prediction using a model based on haplotype clusters. In *ADSA-ASAS Midwest Annual Meeting*, pages 16–16, Des Moines, IA, March 2016. doi: 10.2527/msasas2016-034.

[30] S.D. Kachman, M.L. Spangler, G.L. Bennett, K.J. Hanford, L.A. Kuehn, W.M. Snelling, R.M. Thallman, M. Saatchi, D.J. Garrick, R.D. Schnabel, J.F. Taylor, and E.J. Pollak. Comparison of molecular breeding values based on within- and across-breed training in beef cattle. *Genetics Selection Evolution*, 45:30, 2013.

[31] M. Kendall and A. Stuart. *The Advanced Theory of Statistics*, volume 1. Macmillam Publishing Co., Inc., 4 edition, 1977. ISBN 0028476301.

[32] B.W. Kennedy and L.R. Schaeffer. Genetic evaluation under an animal model when identical genotypes are represented in the population. *Journal of Animal Science*, 67(8):1946–1955, 1989.

[33] L.A. Kuehn, D.J. Nonneman, J.M. Klindt, and T.H. Wise. Genetic relationships of body composition, serum leptin, and age at puberty in gilts. *Journal of animal science*, 87(2):477–483, 2009.

[34] A. Legarra, I. Aguilar, and I. Misztal. A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, 92(9): 4656–4663, 2009.

[35] K.L. Lucot. Genomic predictions for age at puberty and reproductive longevity in sows using Bayesian methods. Master's thesis, University of Nebraska-Lincoln, Lincoln, Nebraska, 2014.

[36] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, 2010.

[37] T.H. Meuwissen and M.E. Goddard. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution*, 36(3):261–279, 2004.

[38] T.H. Meuwissen, B.J. Hayes, and M.E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4): 1819–1829, 2001.

[39] I. Misztal, A. Legarra, and I. Aguilar. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science*, 92(9):4648–4655, 2009.

[40] J.S. Morris. The BLUPs are not "best" when it comes to bootstrapping. *Statistics & Probability Letters*, 56(4):425–430, 2002.

[41] F.D.N. Mujibi, J.D. Nkrumah, O.N. Durunna, P. Stothard, J. Mah, Z. Wang, J. Basarab, G. Plastow, D.H. Crews, and S.S. Moore. Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle. *Journal of Animal Science*, 89(11):3353–3361, 2011.

[42] D.J. Nonneman, J.F. Schneider, C.A. Lents, R.T. Wiedmann, J.L. Vallet, and G.A. Rohrer. Genome-wide association and identification of candidate genes for age at puberty in swine. *BMC genetics*, 17(1):50, 2016.

[43] T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[44] J.L. Patterson, E. Beltranena, and G.R. Foxcroft. The effect of gilt age at first estrus and breeding on third estrus on sow body weight changes and long-term reproductive performance. *Journal of animal science*, 88 (7):2500–2513, 2010.

[45] Y. Pei, J. Li, L. Zhang, C. Papasian, and H. Deng. Analyses and comparison of accuracy of different genotype imputation methods. *PLoS One*, 3(10), 2008. doi: 10.1371/journal.pone.0003551.

[46] P. Pérez and G. de los Campos. Genome-wide regression & prediction with the BGLR statistical package. *Genetics*, 2014. doi: 10.1534/genetics.114.164442.

[47] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org/`.

[48] K. Rønningen. Some properties of the selection index derived by "Henderson's mixed model method". *Zeitschrift für Tierzüchtung und Züchtungsbiologie*, 88(1–4):186–193, 1971.

[49] G.K. Robinson. That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32, 1991.

[50] M.M. Rolf, D.J. Garrick, T. Fountain, H.R. Ramey, R.L. Weaber, J.E. Decker, E.J. Pollak, R.D. Schnabel, and J.F. Taylor. Comparison of Bayesian models to estimate direct genomic values in multi-breed commercial beef cattle. *Genetics Selection Evolution*, 47(1):1, 2015.

[51] M. Saatchi, M.C. McClure, S.D. McKay, M.M. Rolf, J.E. Kim, J. Decker, T.M Taxis, R.H. Chapple, H.R. Ramey, S.L. Northcutt,

S. Bauch, B. Woodward, J.C. Dekkers, R.L. Fernando, D.J. Schnabel, R.D. Garrick, and J.F. Taylor. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution*, 43(40), 2011. doi: 10.1186/1298-9686-43-40.

[52] M. Saatchi, R.D. Schnabel, J.F. Taylor, and D.J. Garrick. Large-effect pleiotropic or closely linked QTL segregate within and across ten US cattle breeds. *BMC genomics*, 15(1):442, 2014.

[53] H. Saito, Y. Sasaki, and Y. Koketsu. Associations between age of gilts at first mating and lifetime performance or culling risk in commercial herds. *Journal of Veterinary Medical Science*, 73(5):555–559, 2011.

[54] M. Sargolzaei and F.S. Schenkel. QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, 25(5):680–681, 2009.

[55] SAS Institute Inc. *SAS/STAT Software, Version 9.4.* Cary, NC, 2002-2012. URL `http://www.sas.com/`.

[56] P. Scheet and M. Stephens. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629–644, 2006.

[57] C. Schlötterer. The evolution of molecular markers-just a matter of fashion? *Nature Reviews Genetics*, 5(1):63–69, 2004.

[58] S.R. Searle, G. Casella, and C.E. McCulloch. *Variance Components.* John Wiley & Sons, 2009.

[59] W.M. Snelling, R. Chiu, J.E. Schein, M. Hobbs, C.A. Abbey, D.L. Adelson, J. Aerts, G.L. Bennett, I.E. Bosdet, M. Boussaha, R. Brauning, A.R. Caetano, M.M. Costa, A.M Crawford, B.P. Dalrymple, A. Eggen, A. Everts-van der Wind, S. Floriot, M. Gautier, C.A. Gill, R.D. Green, R. Holt, O. Jann, S.J.M. Jones, S.M. Kappes, J.W. Keele, P.J. de Jong, D.M. Larkin, H.A. Levin, J.C. McEwan, S. McKay, M.A. Marra, C.A. Mathewson, L.K. Matukumalli, S.S. Moore, B. Murdoch, F.W. Nicholas, K. Osoegawa, A. Roy, H. Salih, L. Schibler, R.D Schnabel, L. Silveri, L.C. Skow, T.P.L. Smith, T.S. Sonstagard, J.F. Talyor, R. Tellam, C.P. Van Tassell, J.L. Williams, J.E. Womack, N.H. Wye, G. Yang, and S. Zhao for the International Bovine BAC Mapping Consortium. A physical map of the bovine genome. *Genome Biology*, 8(8):R165, 2007.

[60] B.E. Stranger, E.A. Stahl, and T. Raj. Review: Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383, 2011.

[61] R.L. Stratonovich. Conditional Markov processes. *Theory of Probability & Its Applications*, 5(2):156–178, 1960.

[62] X. Sun, R.L. Fernando, D.J. Garrick, and J.C.M Dekkers. Improved accuracy of genomic prediction for traits with rare QTL by fitting haplotypes. In *10th World Congress on Genetics Applied to Livestock Production*. ASAS, 2014.

[63] M.D. Trenhaile, J.L. Petersen, S.D. Kachman, R.K. Johnson, and D.C. Ciobanu. Long-term selection for litter size in swine results in shifts in allelic frequency in regions involved in reproductive processes. *Animal genetics*, 47(5):534–542, 2016.

[64] P.M. VanRaden. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423, 2008.

[65] K.A. Weigel, C.P. Van Tassell, J.R. O'Connell, P.M. VanRaden, and G.R. Wiggans. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. *Journal of Dairy Science*, 93(5):2229–2238, 2010.

[66] G. Yang, J. Ren, S. Li, H. Mao, Y. Guo, Z. Zou, D. Ren, J. Ma, and L. Huang. Genome-wide identification of QTL for age at puberty in gilts using a large intercross F2 population between White Duroc x Erhualian. *Genet. Sel. Evol*, 40:529–539, 2008.

[67] N. Yi, B.S. Yandell, G.A. Churchill, D.B. Allison, E.J. Eisen, and D. Pomp. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics*, 170(3):1333–1344, 2005.

[68] X. Zhou, P. Carbonetto, and M. Stephens. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, 9(2), 2013. doi: 10.1371/journal.pgen.1003264.

APPENDIX A

ADDITIONAL RESULTS FOR THE SIMMENTAL DATA SET

A.1 Posterior Distributions

A.1.1 REA

Table A.1.1: REA: Prior and Posterior Means (SE) for Variance Components

Model	Genetic Variance	Residual Variance	Heritability	Haplotype Effect Variance
Prior	0.37	0.55	0.4	0.00073
Bayes B	0.276 (0.017)	0.226 (0.009)	0.549 (0.022)	N/A
Bayes C	0.233 (0.027)	0.237 (0.011)	0.495 (0.038)	N/A
Bayes IM 16	0.268 (0.033)	0.223 (0.013)	0.545 (0.043)	0.00057 (0.00009)
Bayes IM 8	0.287 (0.037)	0.214 (0.014)	0.570 (0.046)	0.00064 (0.00010)
Bayes IM Comp 50	0.288 (0.038)	0.214 (0.015)	0.571 (0.047)	0.00060 (0.00010)
Bayes IM Comp	0.298 (0.043)	0.210 (0.016)	0.584 (0.052)	0.00060 (0.00011)

Figure A.1.1: REA: Density Plots for the Variance Components in Bayes B and C

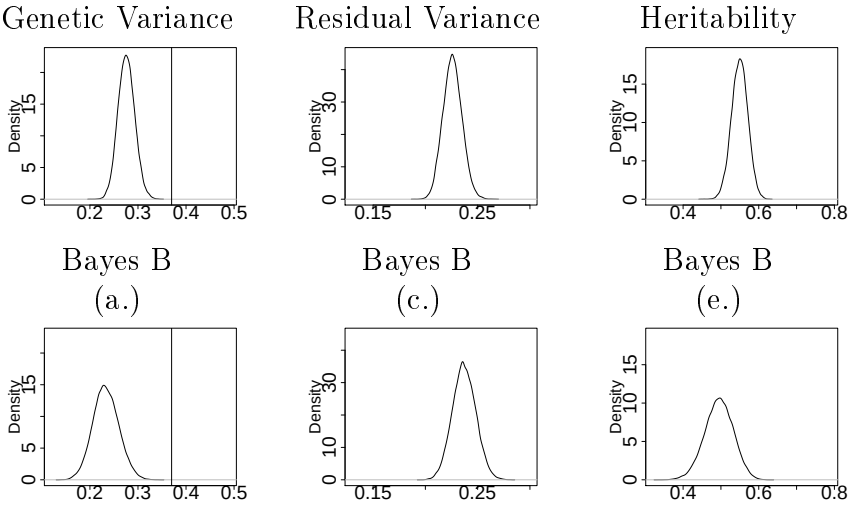
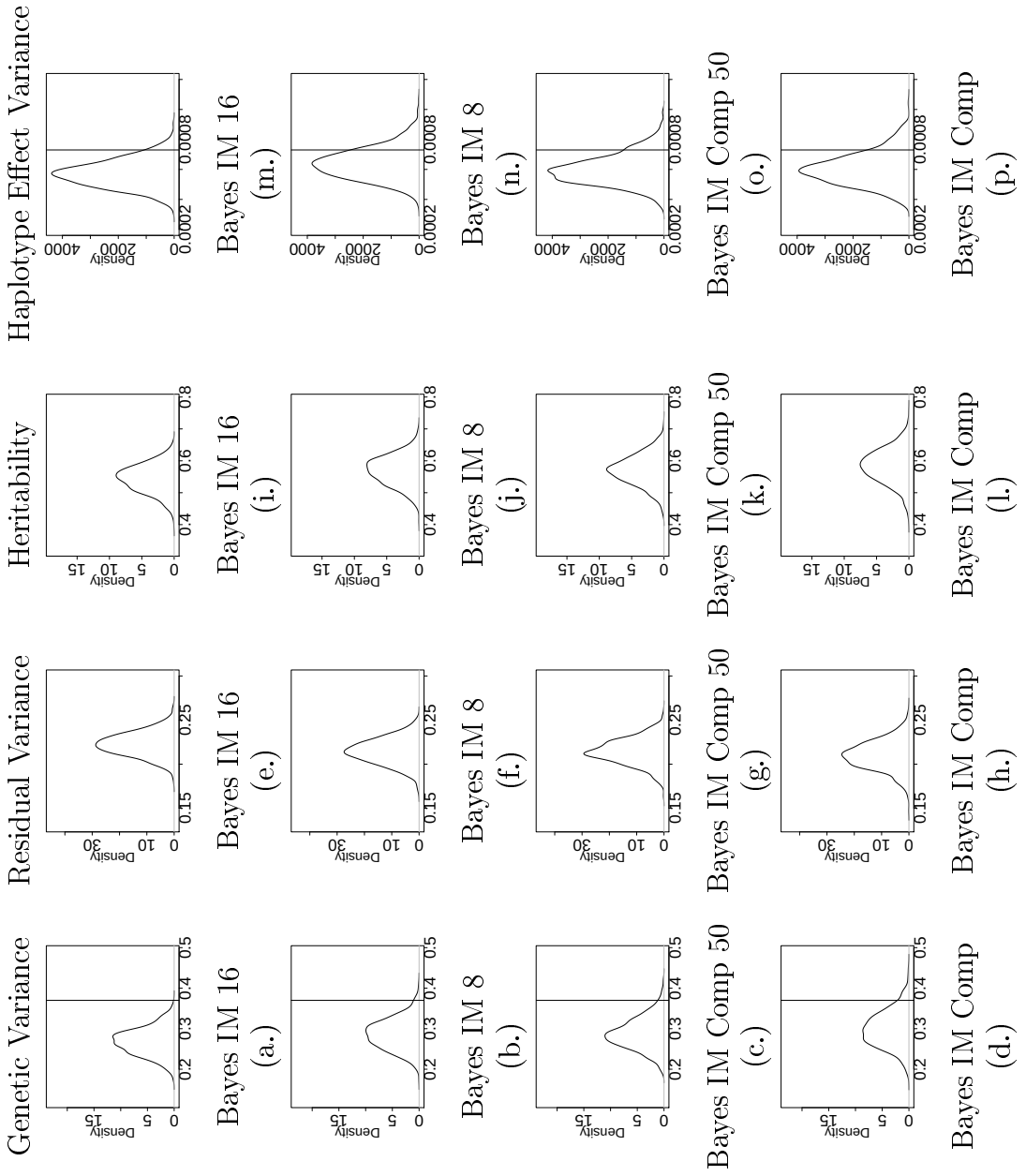


Figure A.1.2: REA: Density Plots for the Variance Components in Bayes IM Models



A.1.2 YG

Table A.1.2: YG: Prior and Posterior Means (SE) for Variance Components

Model	Genetic Variance	Residual Variance	Heritability	Haplotype Effect Variance
Prior	0.077	0.12	0.4	0.00015
Bayes B	0.054 (0.003)	0.058 (0.002)	0.482 (0.020)	N/A
Bayes C	0.040 (0.005)	0.062 (0.003)	0.389 (0.034)	N/A
Bayes IM 16	0.045 (0.006)	0.059 (0.003)	0.431 (0.040)	0.00009 (0.00001)
Bayes IM 8	0.045 (0.006)	0.059 (0.003)	0.433 (0.042)	0.00009 (0.00002)
Bayes IM Comp 50	0.047 (0.007)	0.058 (0.003)	0.444 (0.045)	0.00009 (0.00002)
Bayes IM Comp	0.050 (0.008)	0.057 (0.004)	0.466 (0.053)	0.00009 (0.00002)

Figure A.1.3: YG: Density Plots for the Variance Components in Bayes B and C

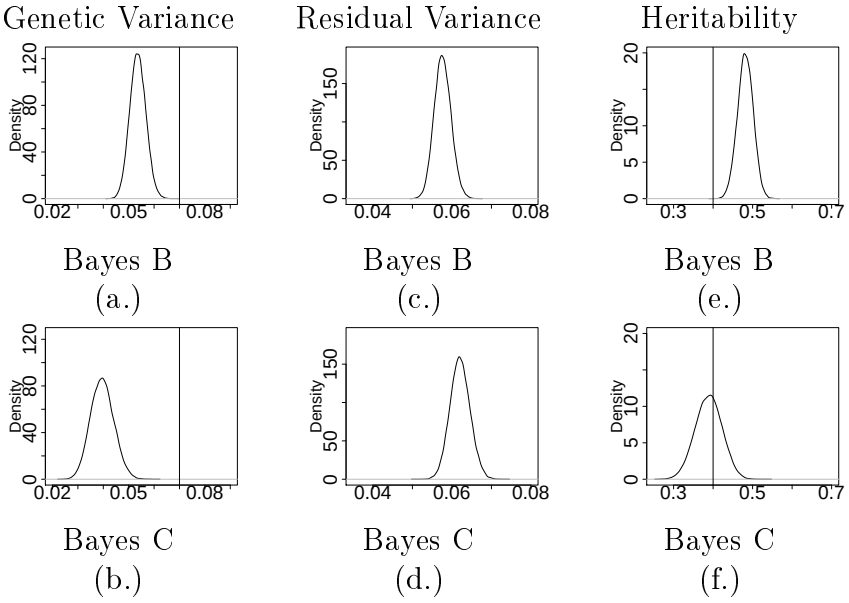
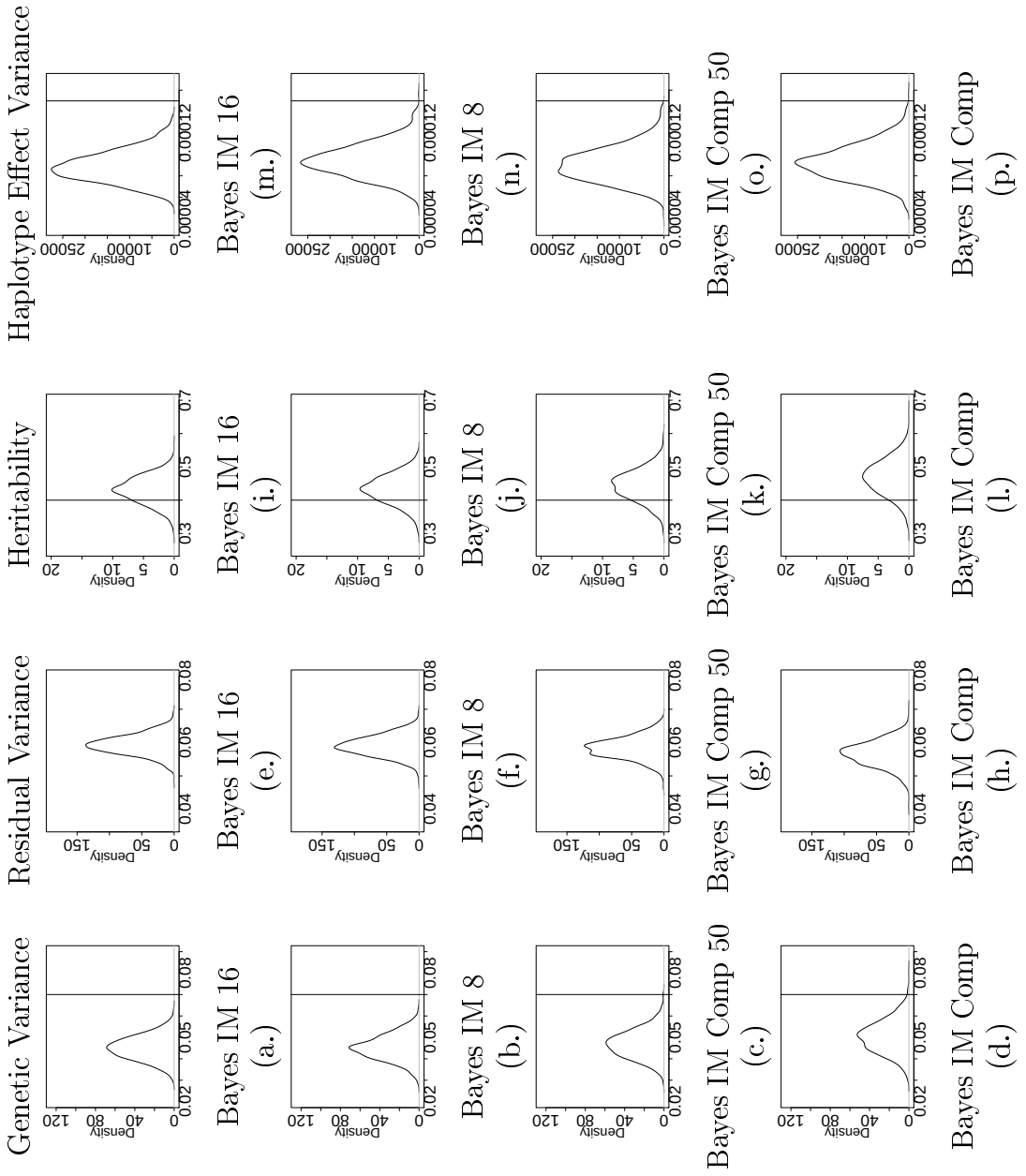


Figure A.1.4: YG: Density Plots for the Variance Components in Bayes IM Models



A.1.3 WWT

Table A.1.3: WWT: Prior and Posterior Means (SE) for Variance Components

Model	Genetic Variance	Residual Variance	Heritability	Haplotype Effect Variance
Prior	243	364	0.4	0.49
Bayes B	224.0 (9.61)	275.0 (9.33)	0.449 (0.016)	N/A
Bayes C	243.4 (12.63)	259.0 (10.29)	0.484 (0.020)	N/A
Bayes IM 16	328.5 (17.02)	186.1 (12.05)	0.638 (0.025)	0.725 (0.067)
Bayes IM 8	326.0 (16.76)	186.9 (11.55)	0.635 (0.025)	0.751 (0.078)
Bayes IM Comp 50	359.2 (17.39)	159.8 (12.02)	0.692 (0.025)	0.790 (0.076)
Bayes IM Comp	393.0 (17.91)	133.34 (12.34)	0.746 (0.025)	0.873 (0.080)

Figure A.1.5: WWT: Density Plots for the Variance Components in Bayes B and C

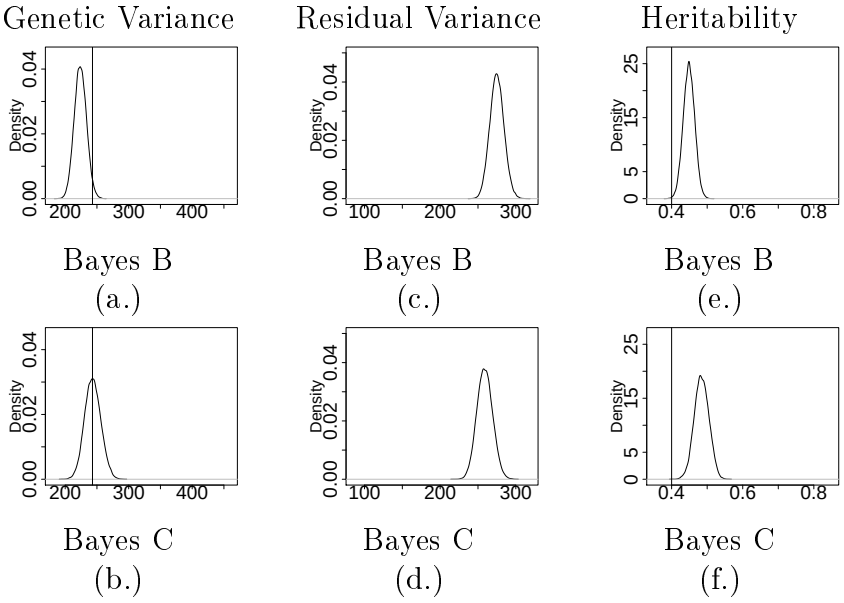
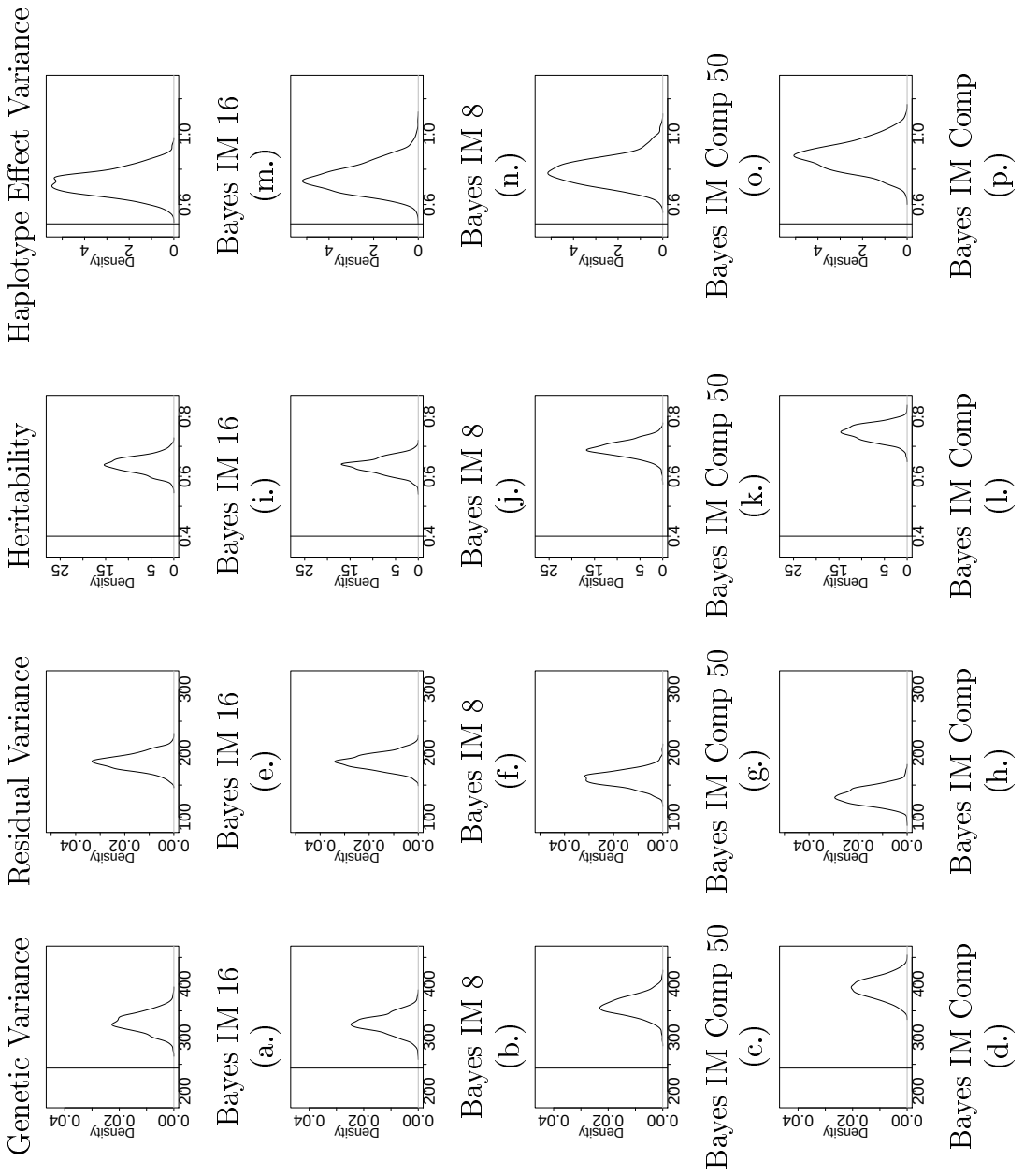


Figure A.1.6: WWT: Density Plots for the Variance Components in Bayes IM Models



A.1.4 YWT

Figure A.1.7: YWT: Density Plots for the Variance Components in Bayes B and C

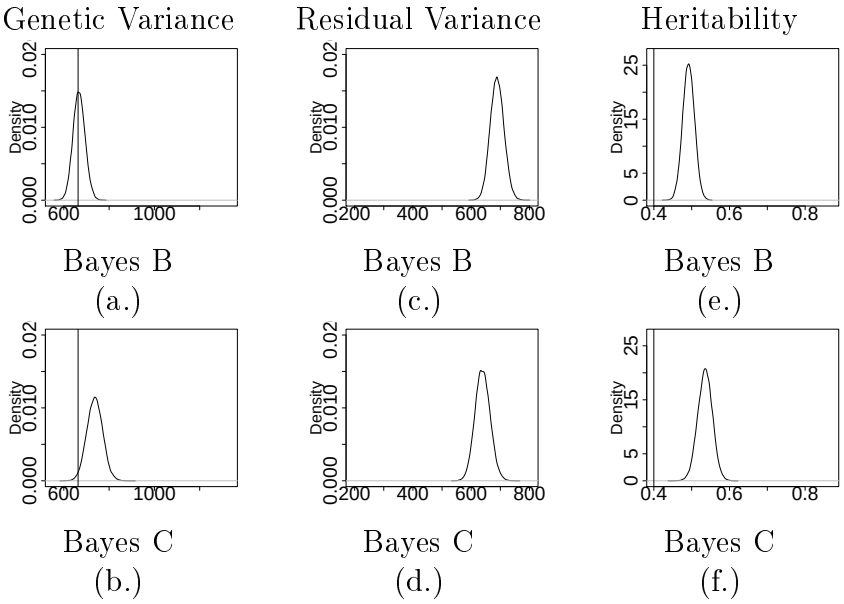
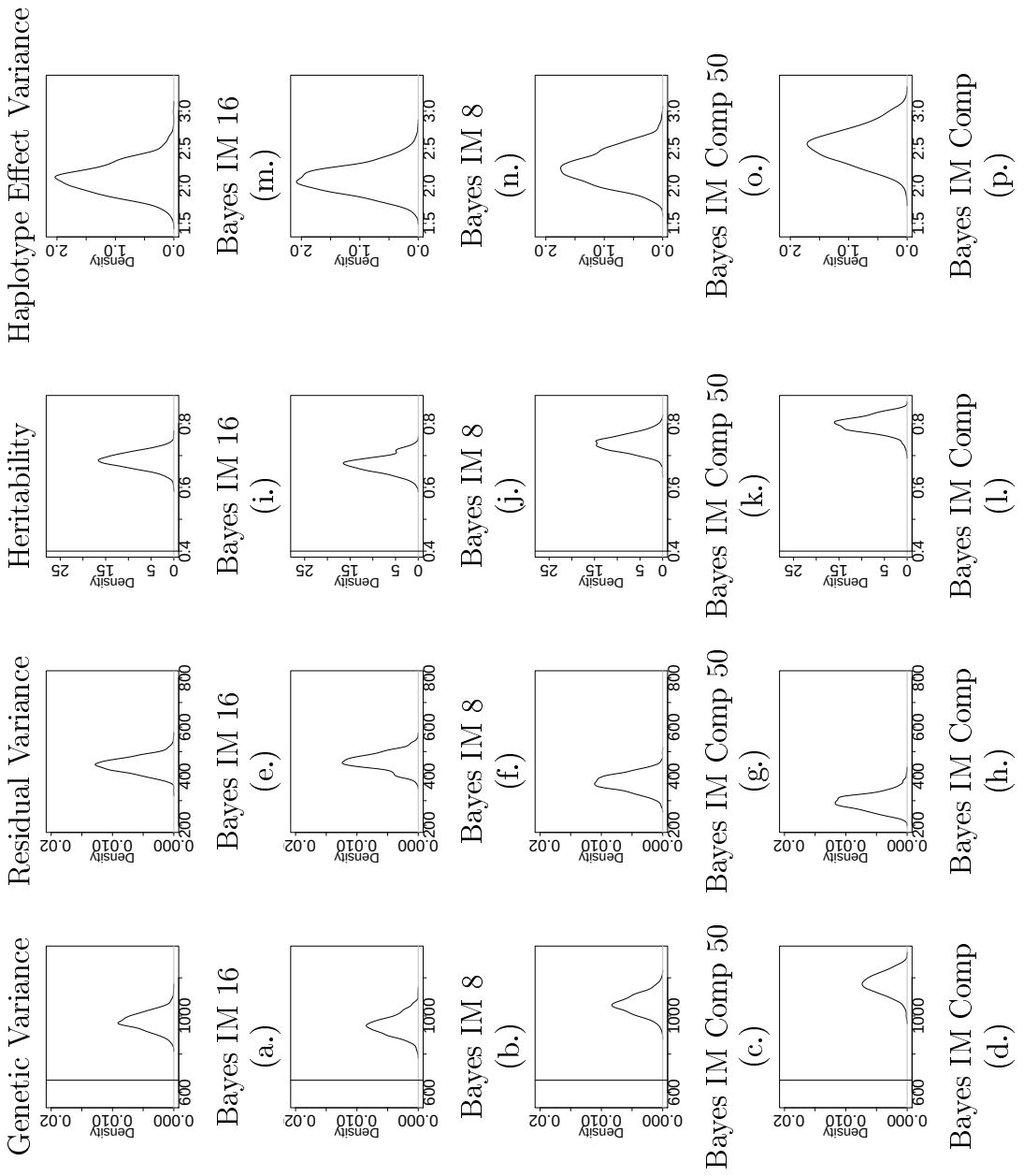


Figure A.1.8: YWT: Density Plots for the Variance Components in Bayes IM Models



A.2 QTL Identification and Haplotype Effects

A.2.1 REA

Figure A.2.1: REA: Genetic Variance for BTA 2 between 2 and 10 MB

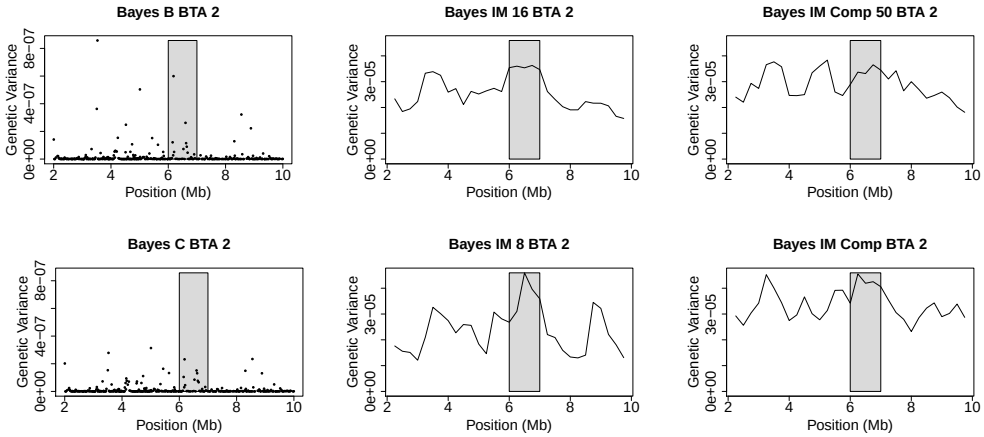


Figure A.2.2: REA: Haplotype Effect Estimates for BTA 2:2 and 10 MB

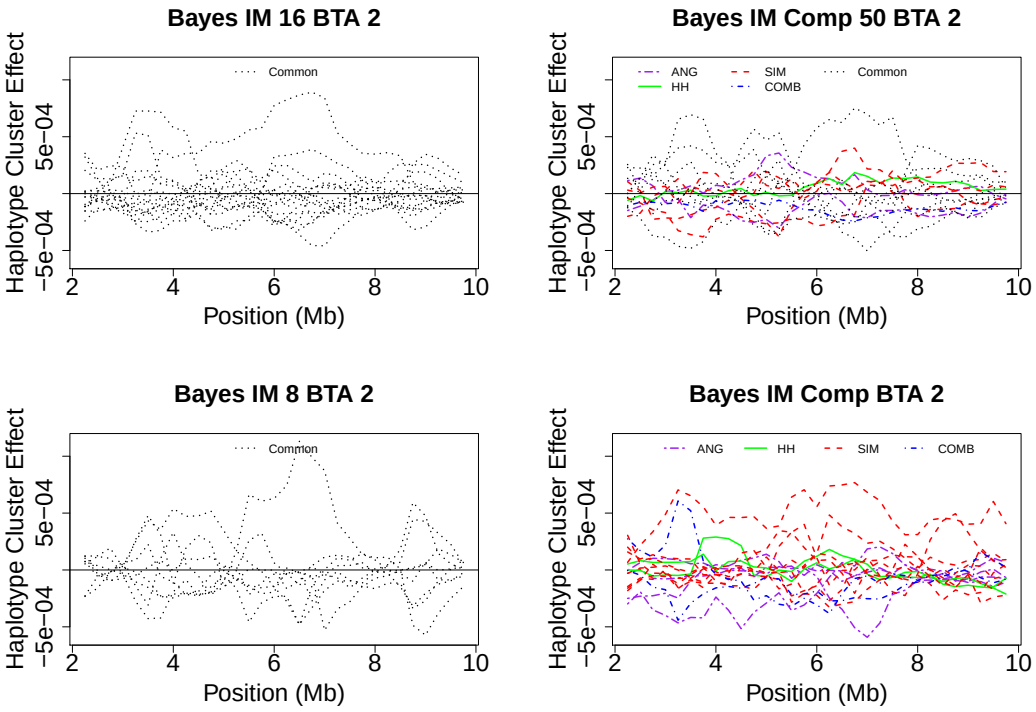


Figure A.2.3: REA: Genetic Variance for BTA 5 between 44 and 52 MB

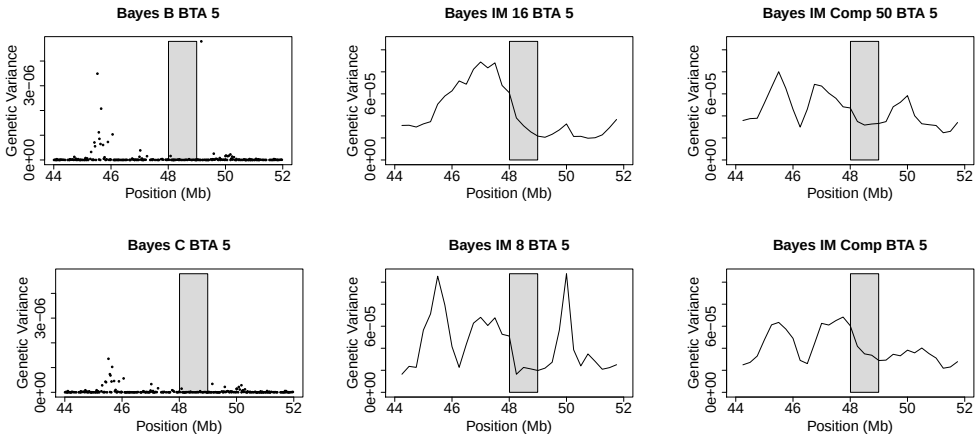


Figure A.2.4: REA: Haplotype Effect Estimates for BTA 5:44 and 52 MB

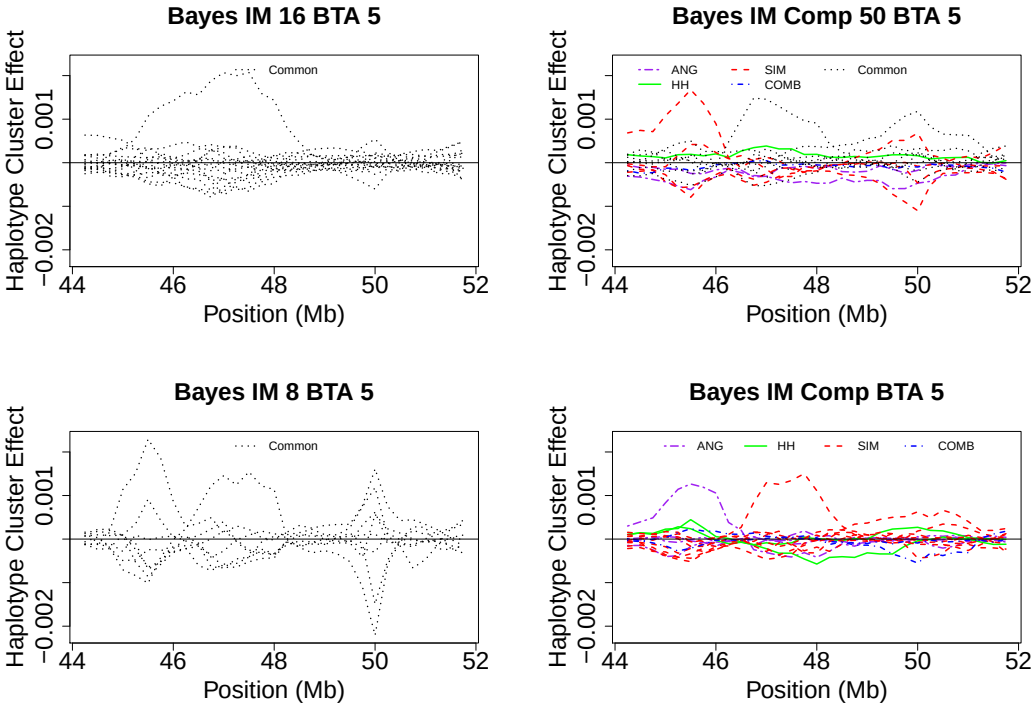


Figure A.2.5: REA: Genetic Variance for BTA 6 between 33 and 43 MB

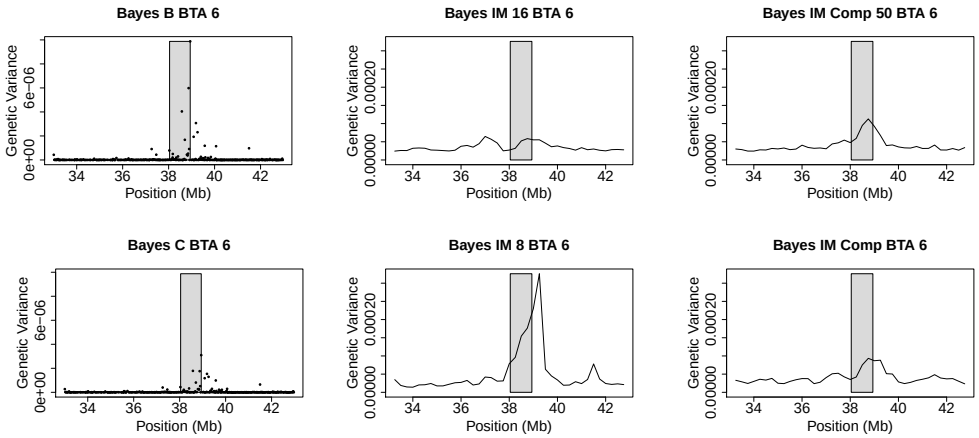


Figure A.2.6: REA: Haplotype Effect Estimates for BTA 6:33 and 43 MB

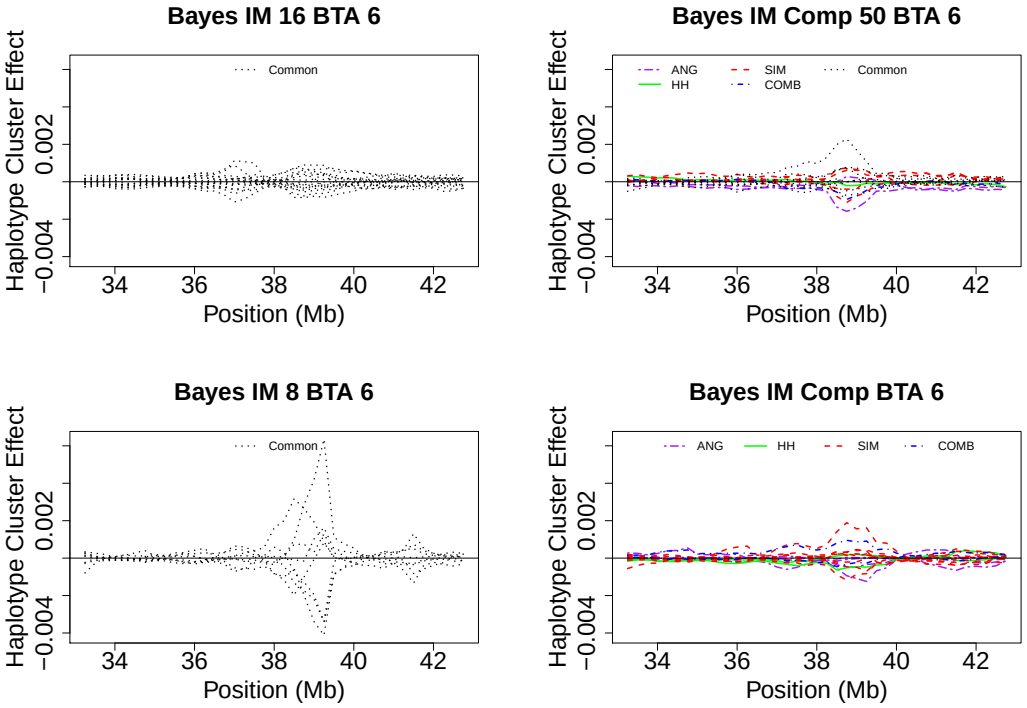


Figure A.2.7: REA: Genetic Variance for BTA 15 between 34 and 42 MB

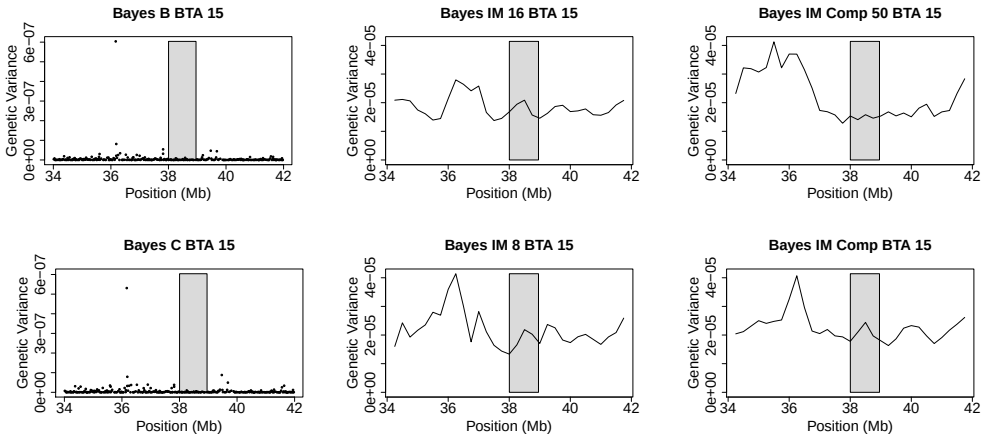
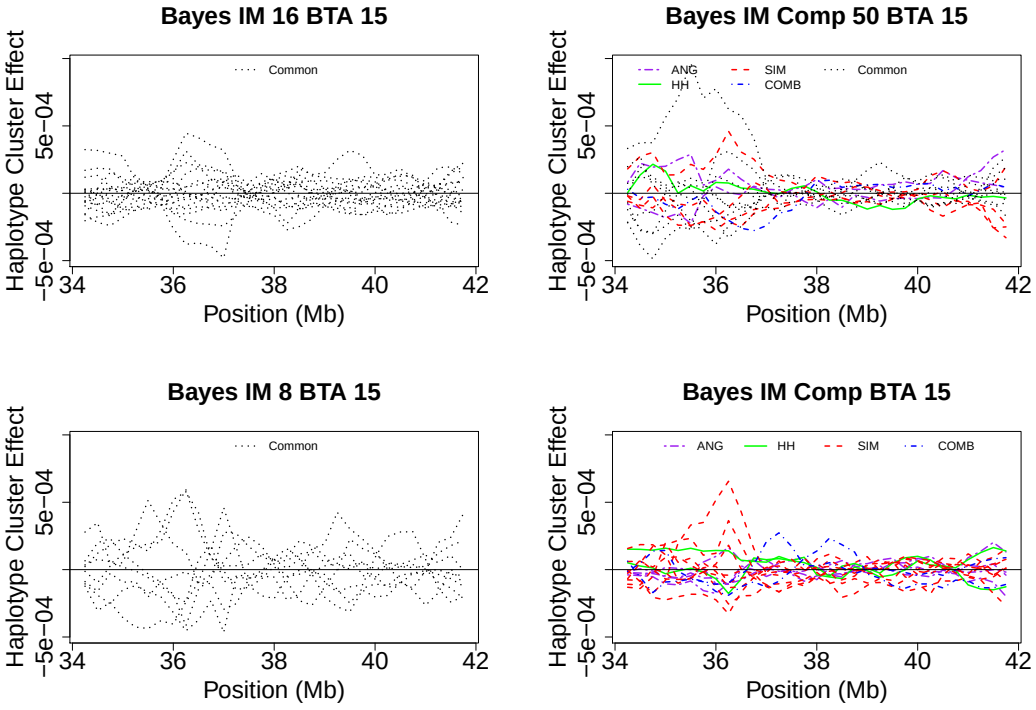


Figure A.2.8: REA: Haplotype Effect Estimates for BTA 15:34 and 42 MB



11**

A.2.2 YG

Figure A.2.9: YG: Genetic Variance for BTA 2 between 2 and 10 MB

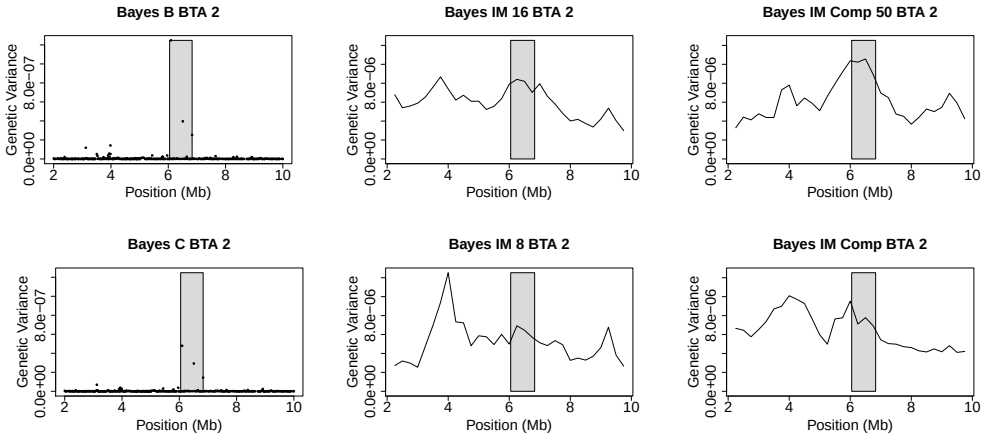


Figure A.2.10: YG: Haplotype Effect Estimates for BTA 2:2 and 10 MB

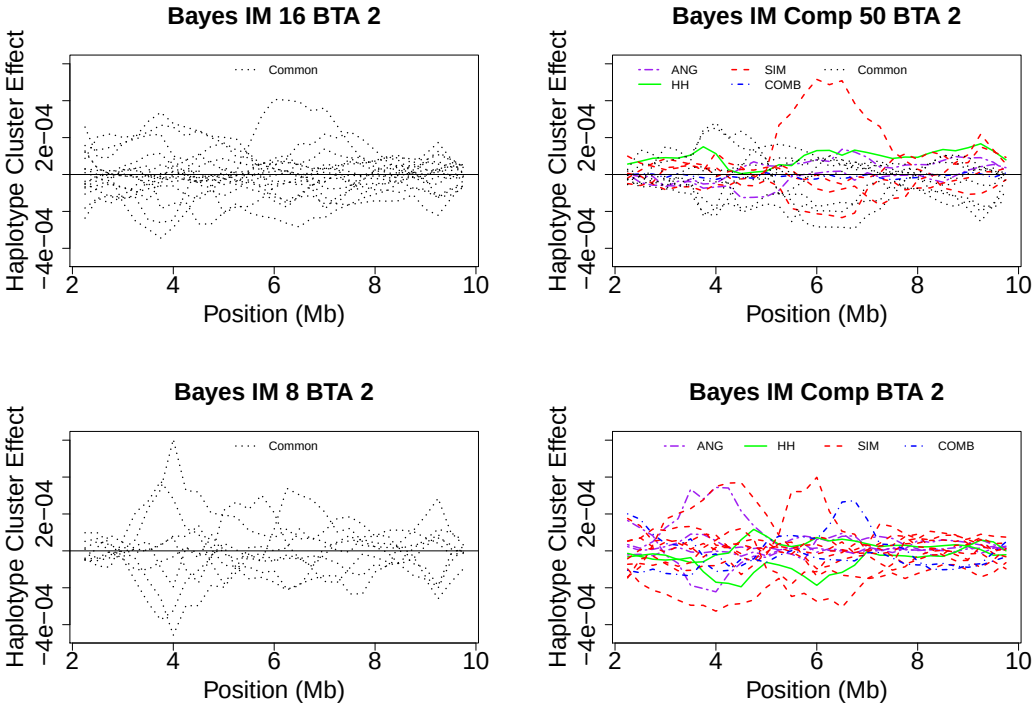


Figure A.2.11: YG: Genetic Variance for BTA 6 between 38 and 46 MB

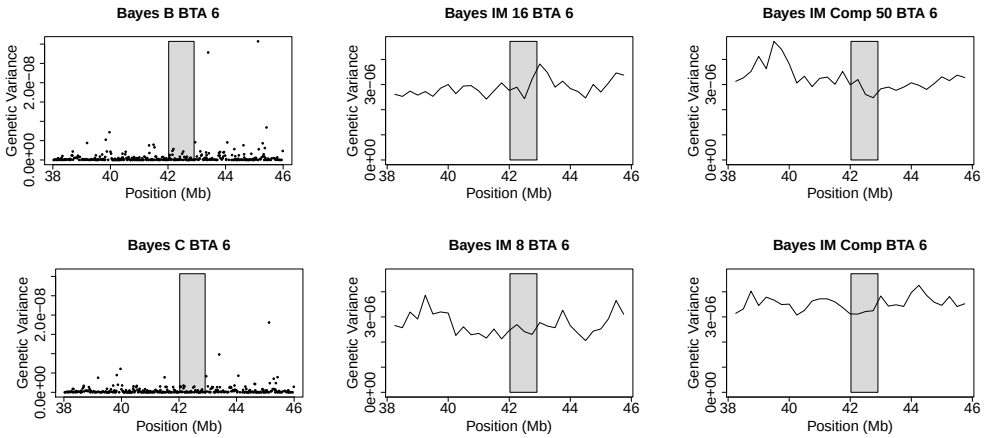
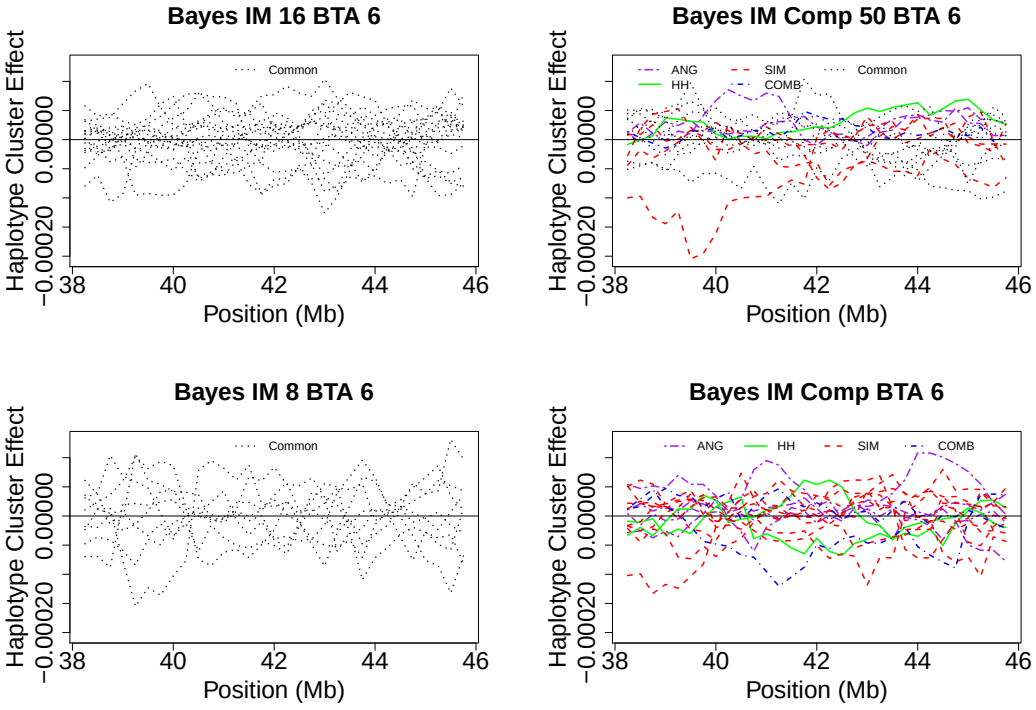


Figure A.2.12: YG: Haplotype Effect Estimates for BTA 6:38 and 46 MB



A.2.3 WWT

Figure A.2.13: WWT: Genetic Variance for BTA 2 between 2 and 10 MB

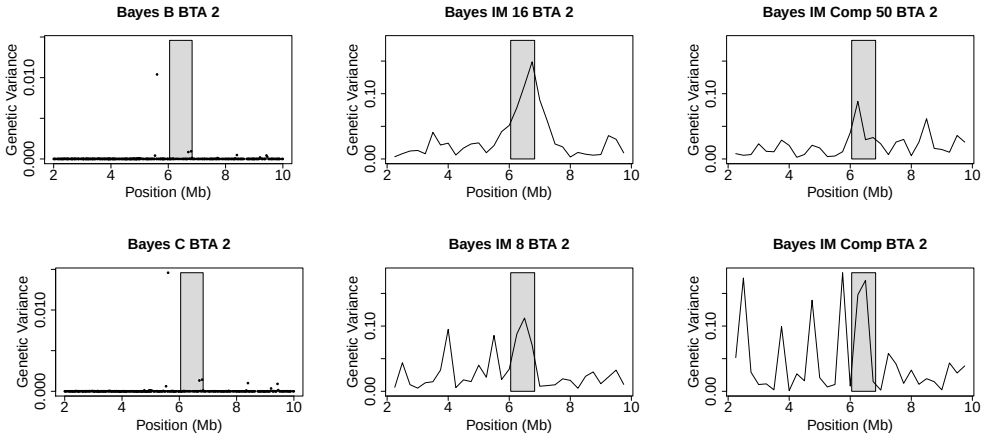


Figure A.2.14: WWT: Haplotype Effect Estimates for BTA 2:2 and 6 MB

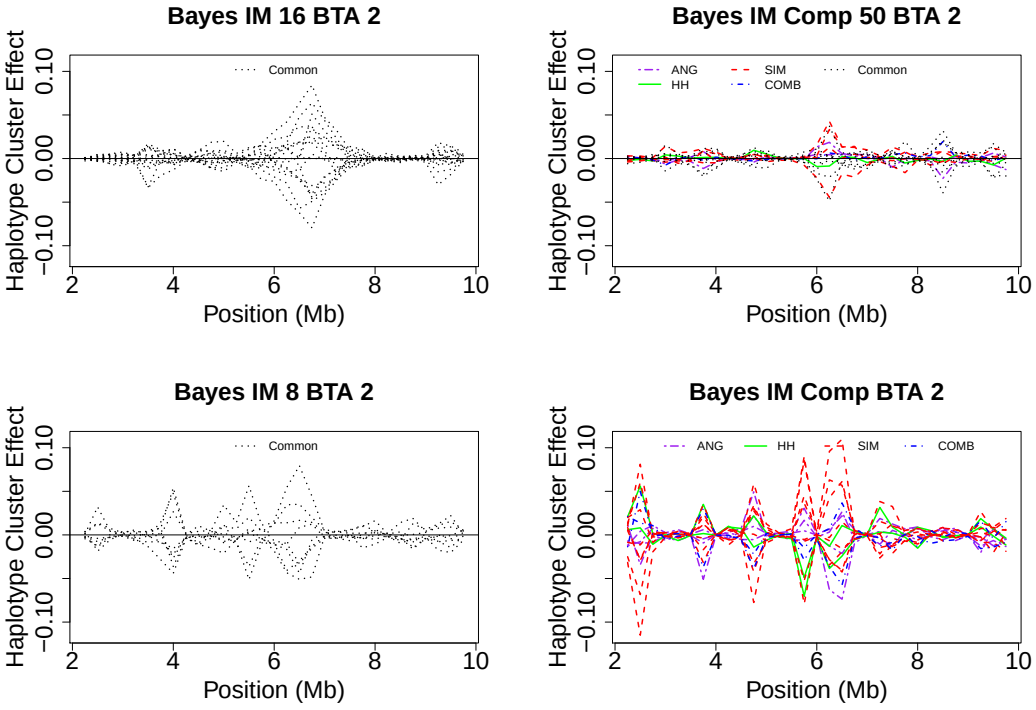


Figure A.2.15: WWT: Genetic Variance for BTA 5 between 102 and 110 MB

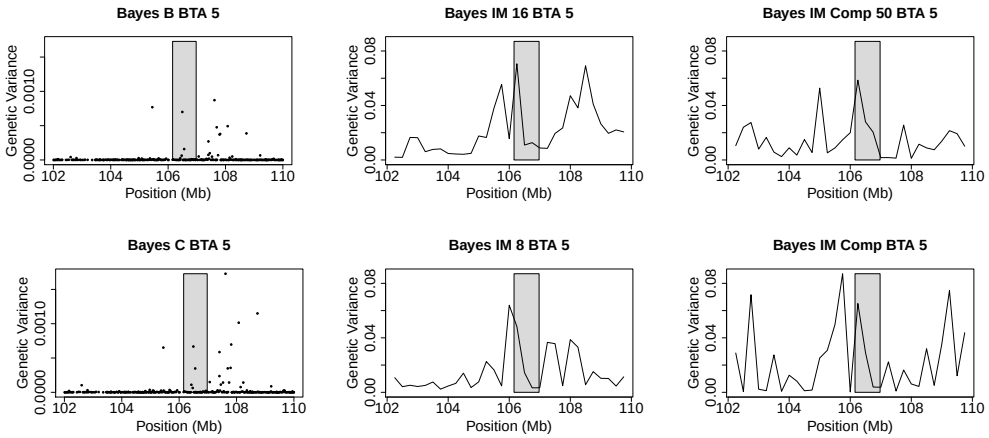


Figure A.2.16: WWT: Haplotype Effect Estimates for BTA 5:102 and 110 MB

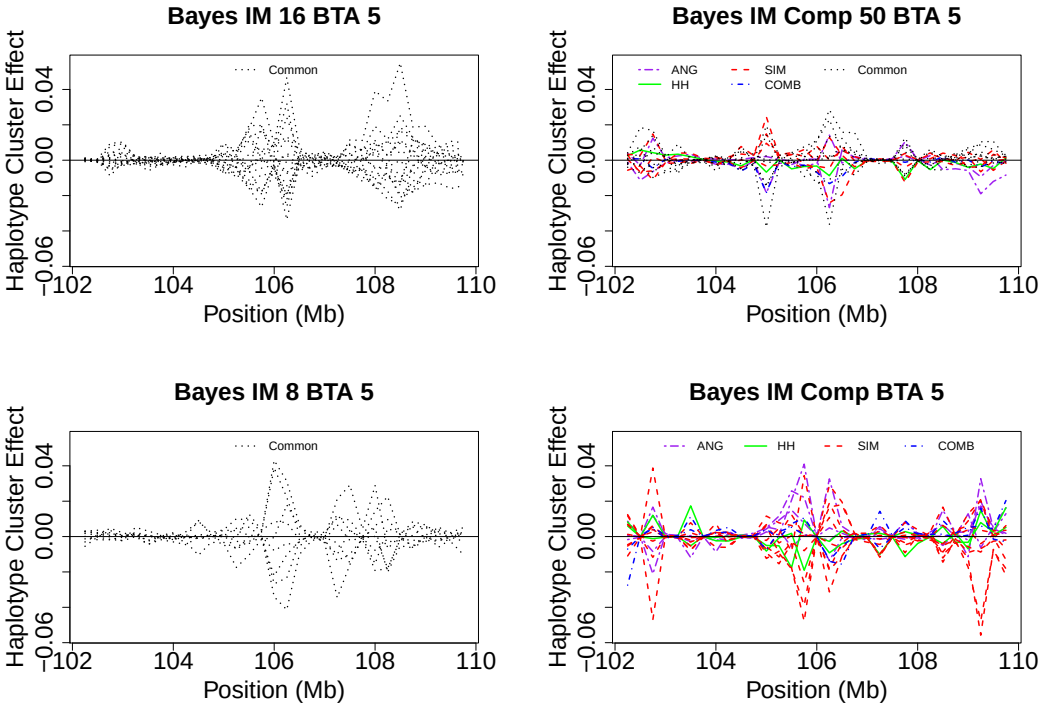


Figure A.2.17: WWT: Genetic Variance for BTA 7 between 89 and 97 MB

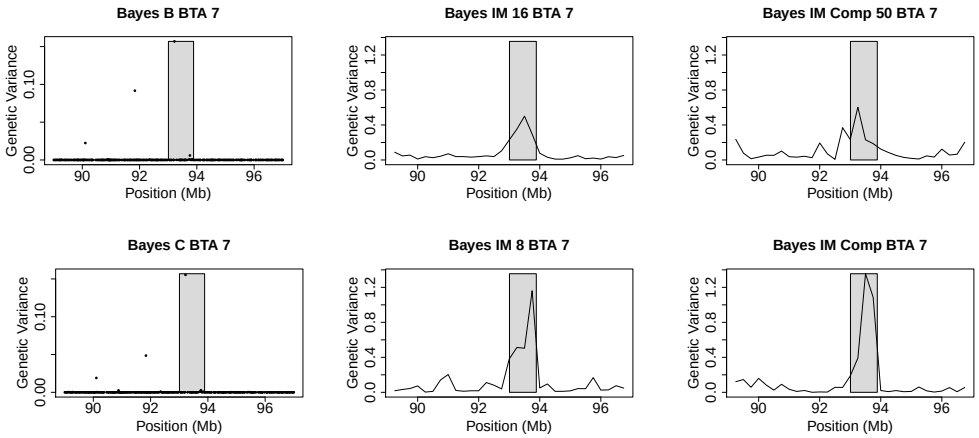


Figure A.2.18: WWT: Haplotype Effect Estimates for BTA 7:89 and 97 MB

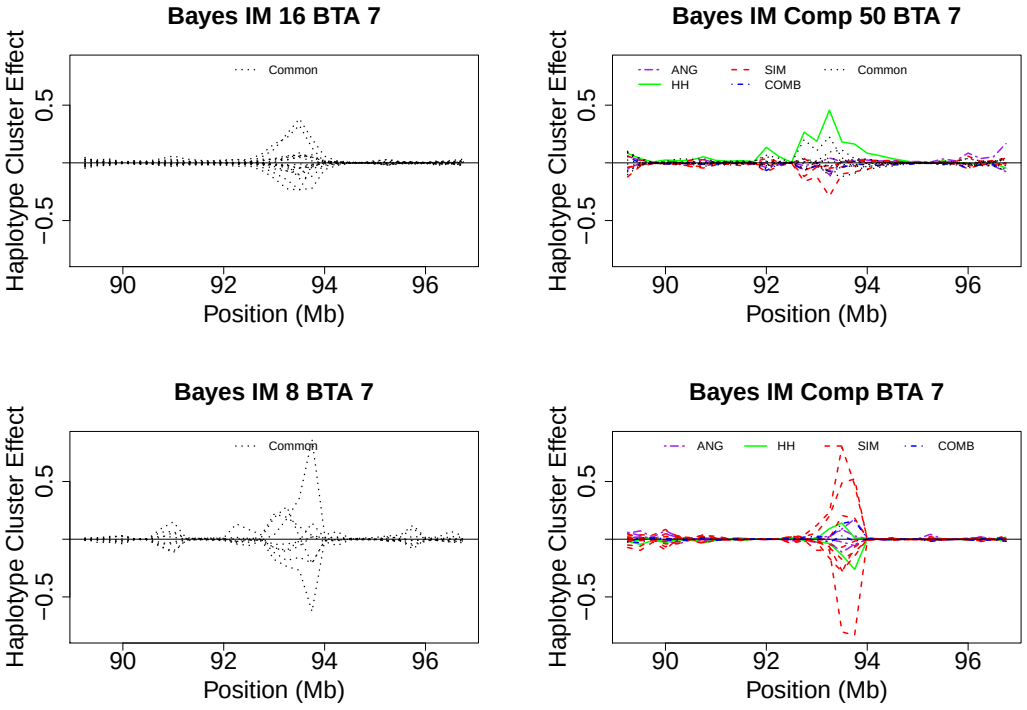


Figure A.2.19: WWT: Genetic Variance for BTA 14 between 19 and 30 MB

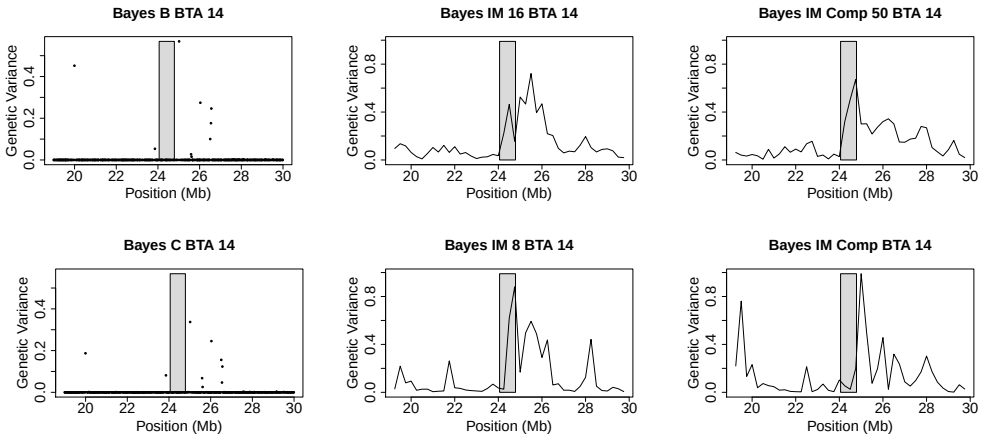


Figure A.2.20: WWT: Haplotype Effect Estimates for BTA 14:19 and 30 MB

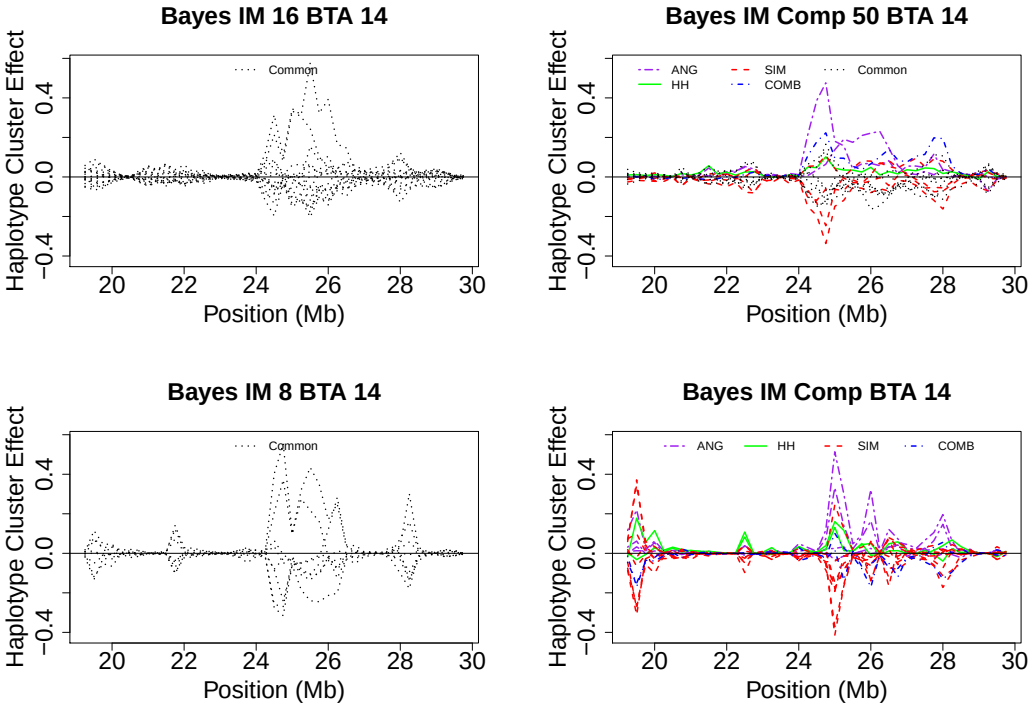


Figure A.2.21: WWT: Genetic Variance for BTA 20 between 0 and 9 MB

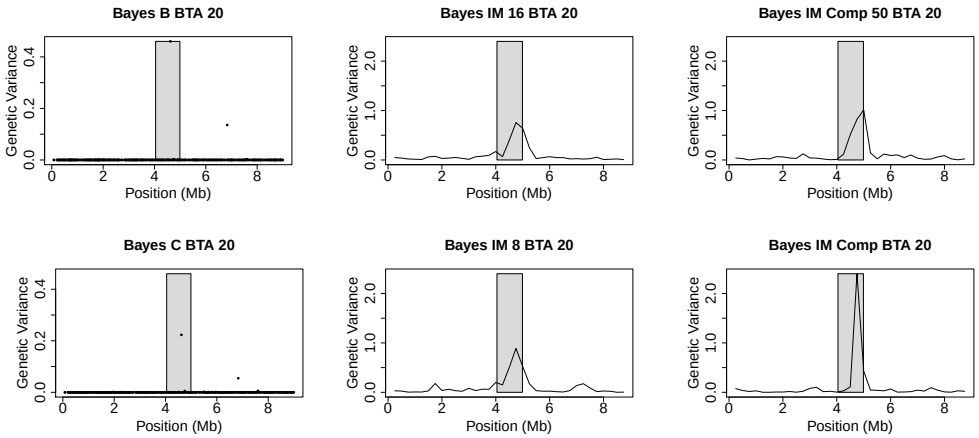
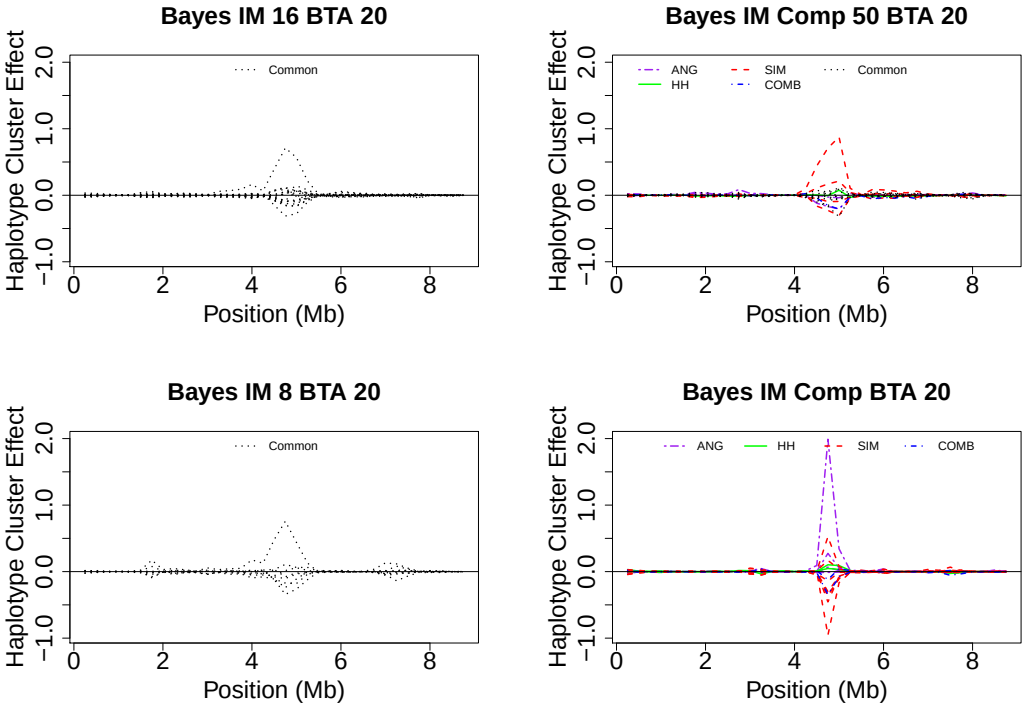


Figure A.2.22: WWT: Haplotype Effect Estimates for BTA 20:0 and 9 MB



A.2.4 YWT

Figure A.2.23: YWT: Genetic Variance for BTA 5 between 102 and 110 MB

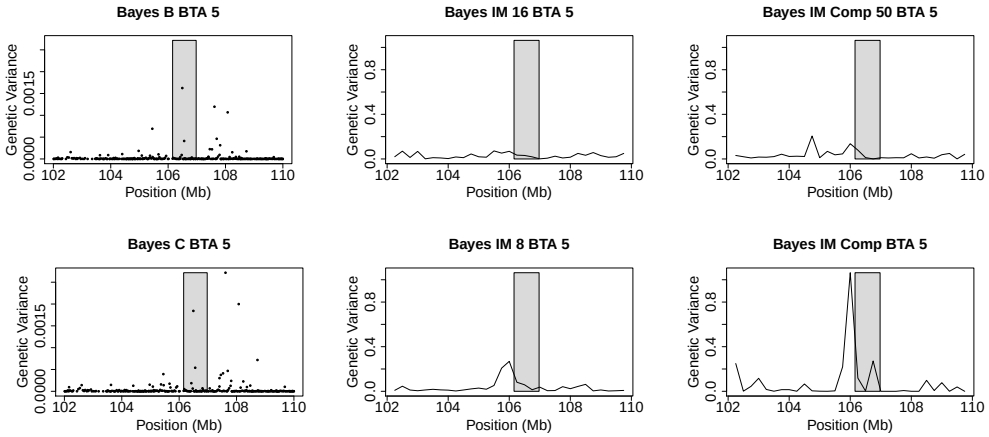


Figure A.2.24: YWT: Haplotype Effect Estimates for BTA 5:102 and 110 MB

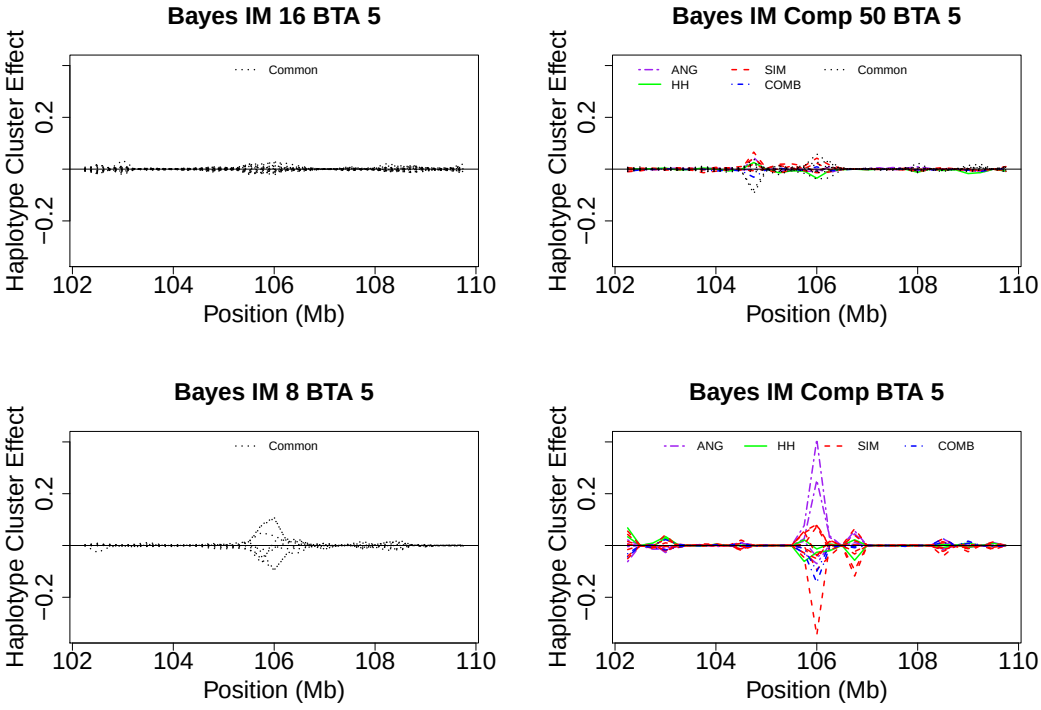


Figure A.2.25: YWT: Genetic Variance for BTA 6 between 33 and 43 MB

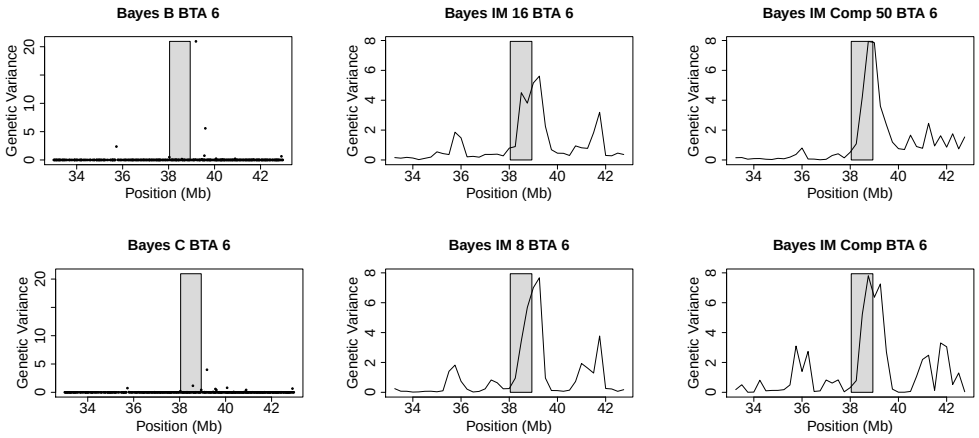


Figure A.2.26: YWT: Haplotype Effect Estimates for BTA 6:33 and 43 MB

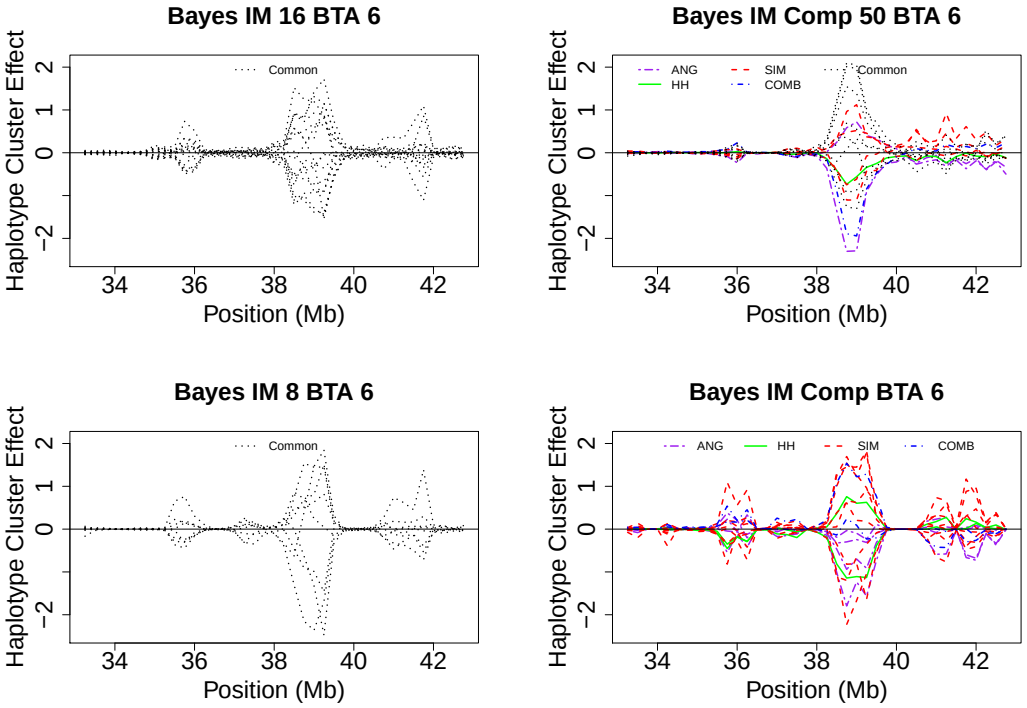


Figure A.2.27: YWT: Genetic Variance for BTA 7 between 89 and 97 MB

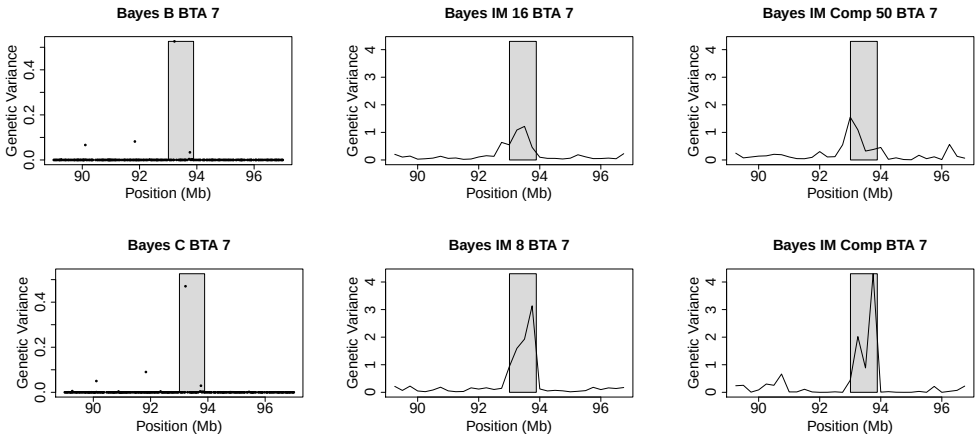


Figure A.2.28: YWT: Haplotype Effect Estimates for BTA 7:89 and 97 MB

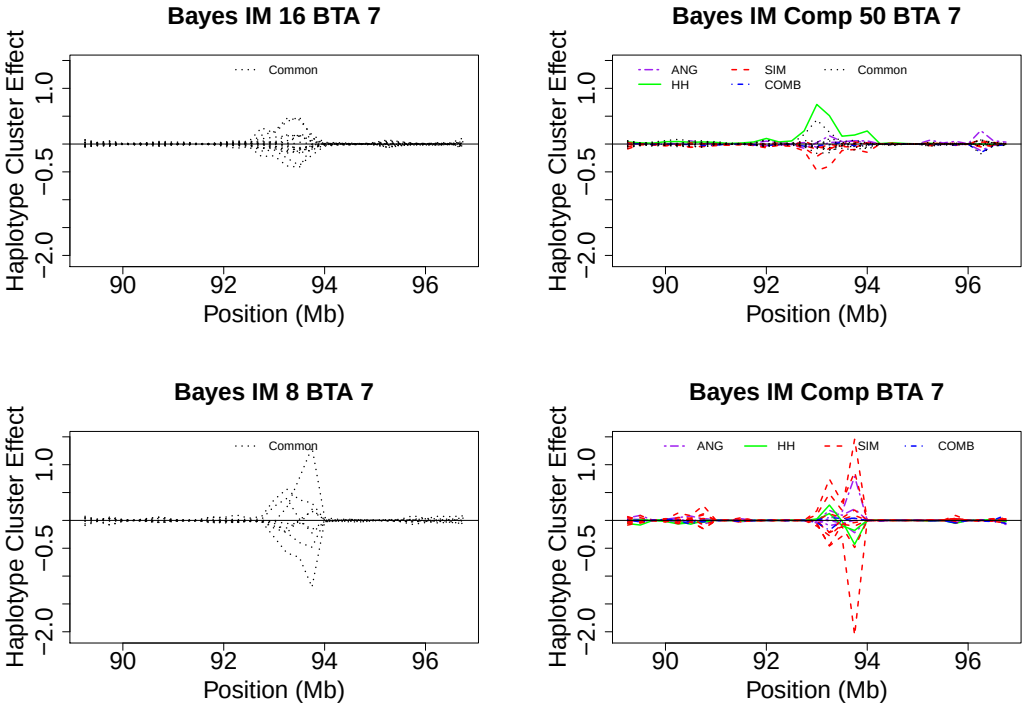


Figure A.2.29: YWT: Genetic Variance for BTA 20 between 0 and 9 MB

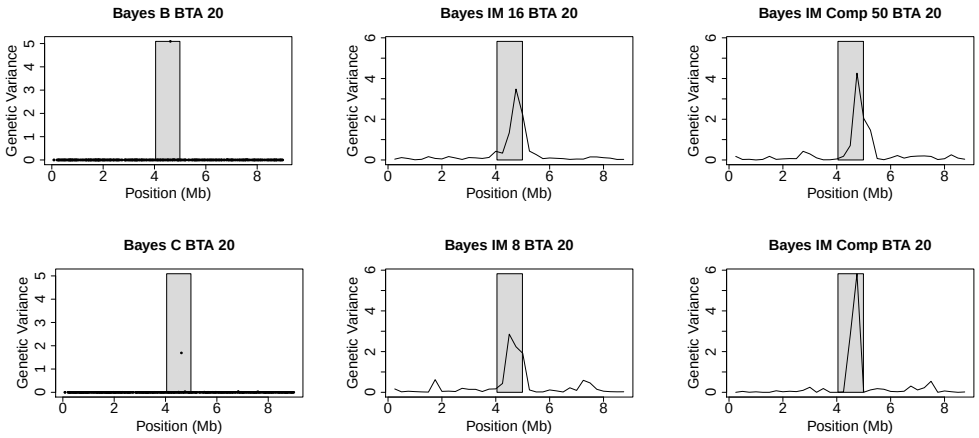
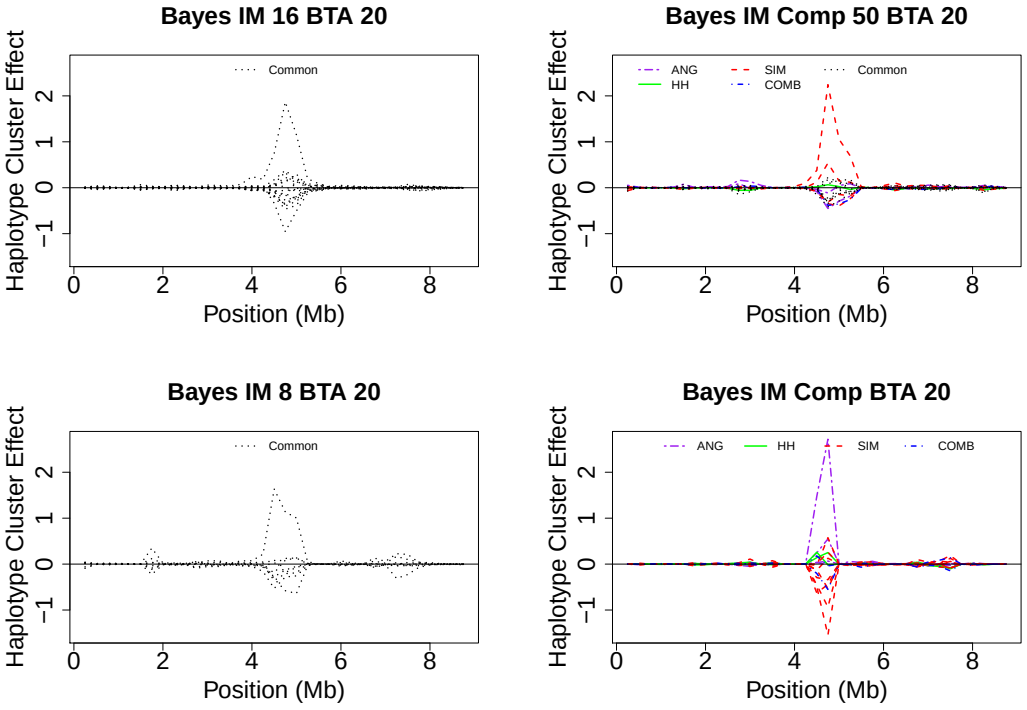


Figure A.2.30: YWT: Haplotype Effect Estimates for BTA 20:0 and 9 MB



A.3 Prediction Accuracy

Table A.3.1: Prediction Accuracy for High Simmental Fold

Traits	Bayes B	Bayes C	Bayes IM 16	Bayes IM 8	Bayes IM Comp 50	Bayes IM Comp
BWT	0.354 (0.034)	0.359 (0.033)	0.354 (0.034)	**0.364 (0.034)**	0.356 (0.034)	0.347 (0.034)
MILK	0.122 (0.050)	0.132 (0.050)	0.156 (0.050)	**0.190 (0.050)**	0.159 (0.050)	0.180 (0.050)
MWWT	0.371 (0.043)	0.366 (0.043)	0.362 (0.043)	0.343 (0.044)	0.358 (0.043)	**0.376 (0.043)**
WWT	0.360 (0.038)	0.360 (0.037)	0.335 (0.038)	0.329 (0.038)	**0.340 (0.038)**	0.330 (0.038)
YWT	0.382 (0.038)	0.393 (0.053)	0.364 (0.038)	0.353 (0.038)	**0.367 (0.038)**	0.365 (0.038)
CWT	0.397 (0.041)	0.396 (0.041)	0.392 (0.041)	**0.410 (0.041)**	0.402 (0.041)	0.390 (0.041)
BFAT	0.265 (0.053)	0.263 (0.053)	**0.233 (0.053)**	0.224 (0.053)	0.218 (0.053)	0.231 (0.053)
MARB	0.366 (0.064)	0.369 (0.064)	0.425 (0.063)	0.447 (0.063)	0.445 (0.062)	**0.458 (0.062)**
REA	0.433 (0.068)	0.432 (0.067)	0.386 (0.070)	0.393 (0.071)	**0.398 (0.070)**	0.383 (0.069)
YG	0.380 (0.062)	0.372 (0.061)	**0.353 (0.061)**	0.340 (0.062)	0.327 (0.061)	0.322 (0.0613)
CE	0.508 (0.034)	0.524 (0.034)	0.511 (0.034)	**0.522 (0.034)**	0.504 (0.034)	0.510 (0.034)
DOC	0.183 (0.055)	0.201 (0.053)	0.188 (0.053)	0.178 (0.054)	**0.191 (0.053)**	0.177 (0.053)
MCE	0.299 (0.046)	0.296 (0.046)	0.271 (0.046)	0.266 (0.047)	0.250 (0.046)	**0.290 (0.046)**

a. Genetic correlation (SE)
b. Best model out of all 6 is underlined
c. Best model out of all Bayes IM models is in bold.

Table A.3.2: Score Differential for the High Simmental Fold

	Model A					
Model B	Bayes B	Bayes C**	Bayes IM 16	Bayes IM 8	Bayes IM Comp 50	Bayes IM Comp
Bayes B		2	-4	-3	-3	-5
Bayes C	-2		-9	-5	-7	-7
Bayes IM 16	4	9		-1	3	-3
Bayes IM 8	3	5	1		-1	-1
Bayes IM Comp 50	3	7	-3	1		-1
Bayes IM Comp	5	7	3	1	1	
Total Score Differential	13	30**	-12	-7	-7	-17

a. Score differential = (# of traits Model A had higher prediction accuracy) - (# of trait Model B had higher prediction accuracy)
b. ** represents the model with the highest total score differential

Table A.3.3: Prediction Accuracy for Medium Simmental Fold

Traits	Bayes B	Bayes C	Bayes IM 16	Bayes IM 8	Bayes IM Comp 50	Bayes IM Comp
BWT	0.563 (0.036)	0.549 (0.036)	0.551 (0.036)	0.547 (0.036)	**0.554 (0.036)**	0.550 (0.036)
MILK	0.206 (0.065)	0.202 (0.065)	0.236 (0.065)	**0.292 (0.065)**	0.255 (0.065)	0.230 (0.064)
MWWT	0.484 (0.055)	0.471 (0.055)	0.564 (0.055)	0.580 (0.054)	0.578 (0.054)	**0.599 (0.053)**
WWT	0.457 (0.043)	0.472 (0.043)	0.527 (0.042)	**0.550 (0.041)**	0.539 (0.041)	0.531 (0.041)
YWT	0.510 (0.043)	0.514 (0.042)	0.563 (0.041)	**0.575 (0.041)**	0.562 (0.041)	0.558 (0.041)
CWT	0.612 (0.044)	0.597 (0.044)	0.632 (0.043)	**0.662 (0.043)**	0.637 (0.043)	0.619 (0.043)
BFAT	0.326 (0.051)	0.356 (0.051)	0.301 (0.052)	0.312 (0.052)	0.317 (0.051)	**0.326 (0.051)**
MARB	0.467 (0.063)	0.462 (0.063)	0.483 (0.062)	0.487 (0.063)	0.467 (0.063)	**0.500 (0.063)**
REA	0.452 (0.073)	0.426 (0.073)	**0.411 (0.074)**	0.397 (0.075)	0.402 (0.074)	0.398 (0.075)
YG	0.475 (0.066)	0.472 (0.065)	**0.403 (0.067)**	0.401 (0.067)	0.390 (0.067)	0.398 (0.066)
CE	0.741 (0.033)	0.738 (0.033)	0.756 (0.032)	0.759 (0.033)	0.751 (0.033)	**0.764 (0.032)**
DOC	0.218 (0.074)	0.317 (0.069)	**0.302 (0.069)**	0.259 (0.070)	0.298 (0.069)	0.286 (0.069)
MCE	0.420 (0.058)	0.392(0.058)	0.421 (0.058)	0.385 (0.059)	0.412 (0.058)	**0.425 (0.058)**

a. Genetic correlation (SE)
b. Best model out of all 6 is underlined
c. Best model out of all Bayes IM models is in bold.

Table A.3.4: Score Differential for the Medium Simmental Fold

	Model A					
Model B	Bayes B	Bayes C	Bayes IM 16	Bayes IM 8	Bayes IM Comp 50	Bayes IM Comp**
Bayes B		-5	4	3	2	6
Bayes C	5		5	1	5	4
Bayes IM 16	-4	-5		3	0	0
Bayes IM 8	-3	-1	-3		-3	2
Bayes IM Comp 50	-2	-5	0	3		-1
Bayes IM Comp	-6	-4	0	-2	1	
Total Score Differential	-10	-20	6	-8	5	11**

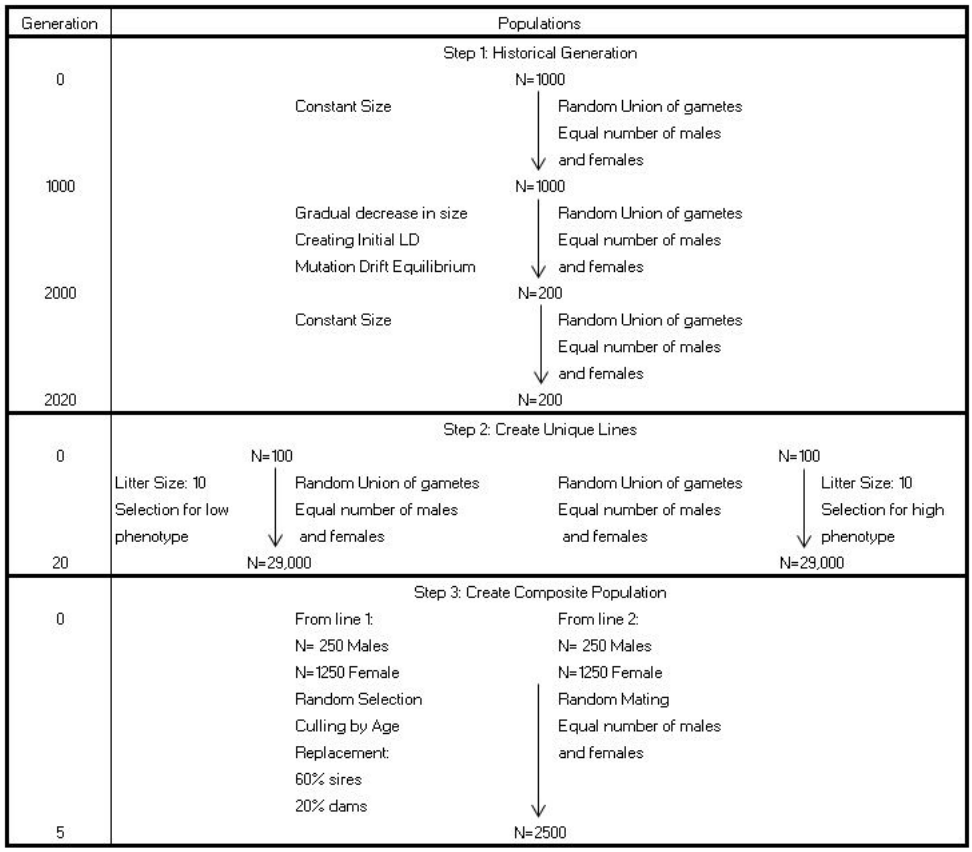
a. Score differential = (# of traits Model A had higher prediction accuracy) - (# of trait Model B had higher prediction accuracy)
b. ** represents the model with the highest total score differential

APPENDIX B

ADDITIONAL RESULTS FOR THE SIMULATION DATA SET

B.1 Data Generation Information

Figure B.1.1: Map of the Population Simulation



a. This figure has been adapted from Brito et al. [6].

Table B.1.1: Summary of SNP and QTL marker information

BTA	Length (MB)	Number of QTL (Active)	Number of SNP (Polymorphic)	Distance Between Polymorphic SNPs					r^2 between adjacent SNPs		
				Min. (Kb)	25% (Kb)	50% (Kb)	75% (Kb)	Max (Kb)	25%	50%	75%
1	146	40 (22)	11,027 (3,733)	10	7,170	18,890	41,7023	876,350	0.133	0.971	0.999
2	126	32 (18)	8,990 (3,317)	10	7,754	19,115	40,910	680,950	0.144	0.975	0.999
3	116	30 (12)	8,367 (2,924)	9	7,855	18,940	41,460	755,110	0.169	0.975	0.999
4	111	44 (14)	8,083 (3,112)	10	7,441	18,070	37,911	668,070	0.177	0.976	0.999
5	119	34 (11)	7,337 (2,693)	40	8,407	22,830	49,125	878,000	0.142	0.901	0.998
6	112	44 (17)	9,663 (3,223)	9	6,073	16,460	35,219	928,020	0.187	0.984	0.999
7	101	32 (11)	8,407 (2,932)	39	6,260	16,390	35,710	816,330	0.161	0.982	0.999
8	104	30 (8)	7,787 (2,522)	10	7,820	18,360	41,571	950,469	0.180	0.985	0.999
9	95	30 (11)	7,213 (2,711)	29	6,840	17,805	38,803	686,049	0.136	0.961	0.999
10	96	36 (17)	7,717 (2,907)	20	6,860	17,160	35,994	632,000	0.167	0.965	0.999
11	102	26 (8)	7,233 (2,748)	40	7,960	19,910	41,615	735,610	0.147	0.962	0.999
12	78	26 (6)	5,773 (1,905)	20	7,843	19,245	42,903	719,249	0.103	0.906	0.999
13	83	38 (17)	5,693 (2,061)	10	8,120	21,045	44,225	714,350	0.190	0.957	0.998
14	82	56 (23)	5,643 (1,904)	50	7,345	19,890	44,476	811,070	0.197	0.979	0.999
15	75	28 (8)	5,623 (1,811)	10	7,573	18,521	41,977	858,270	0.130	0.983	0.999
16	73	46 (14)	5,570 (2,009)	20	7,370	18,240	38,768	784,220	0.105	0.944	0.999
17	70	36 (12)	5,223 (1,902)	10	7,149	17,791	39,330	780,120	0.219	0.982	0.999
18	63	26 (7)	4,403 (1,798)	31	7,530	19,200	39,670	837,851	0.165	0.902	0.998
19	63	38 (13)	4,637 (1,408)	50	7,270	19,560	44,470	950,521	0.073	0.895	0.999
20	68	32 (10)	5,370 (2,027)	9	7,102	17,840	37,435	1,070,950	0.157	0.979	0.999
21	63	32 (10)	4,837 (1,642)	30	6,919	17,280	39,150	850,050	0.183	0.978	0.999
22	60	38 (12)	4,040 (1,616)	10	7,650	19,220	43,025	439,400	0.156	0.951	0.999
23	49	22 (9)	3,753 (1,434)	10	7,240	17,661	38,990	716,911	0.142	0.870	0.998
24	60	32 (11)	4,077 (1,420)	51	7,960	20,341	46,525	737,550	0.178	0.956	0.999
25	42	44 (17)	3,063 (1,025)	69	7,555	20,286	44,369	1,345,750	0.192	0.976	0.999
26	48	40 (13)	3,480 (1,268)	10	7,560	18,911	41,860	708,640	0.168	0.980	0.999
27	43	30 (6)	3,117 (1,047)	21	7,367	18,005	38,608	794,060	0.214	0.985	0.999
28	40	28 (13)	3,013 (1,126)	60	7,149	18,029	38,800	868,590	0.168	0.978	0.999
29	45	30 (11)	3,393 (1,130)	20	7,511	19,120	44,170	877,490	0.165	0.984	0.999
Overall	**2333**	**1,000 (361)**	**172,532 (61,255)**	**9**	**7,330**	**18,600**	**40,640**	**1,345,750**	**0.155**	**0.971**	**0.999**

B.2 Posterior Distributions

Figure B.2.1: Density Plots for the Variance Components in Bayes B and C

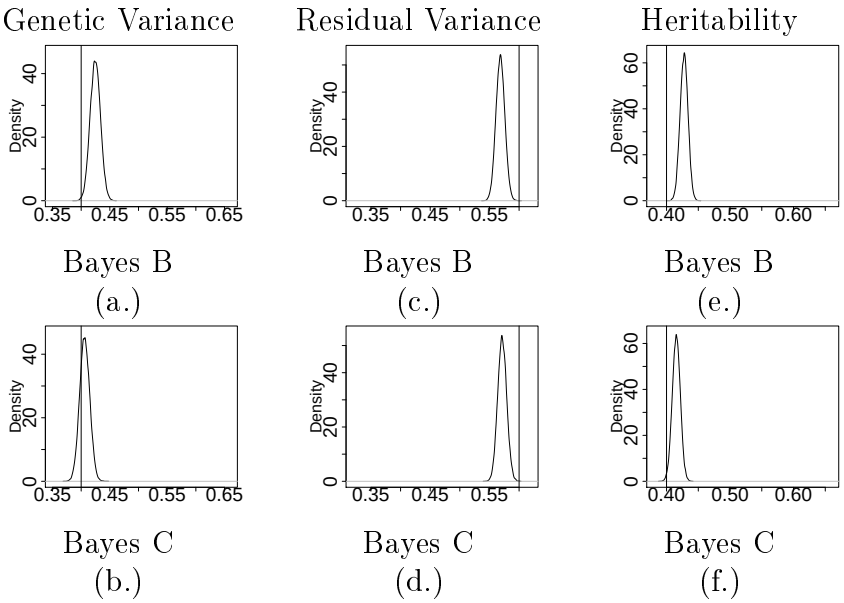


Figure B.2.2: Density Plots for the Variance Components in Bayes IM Models

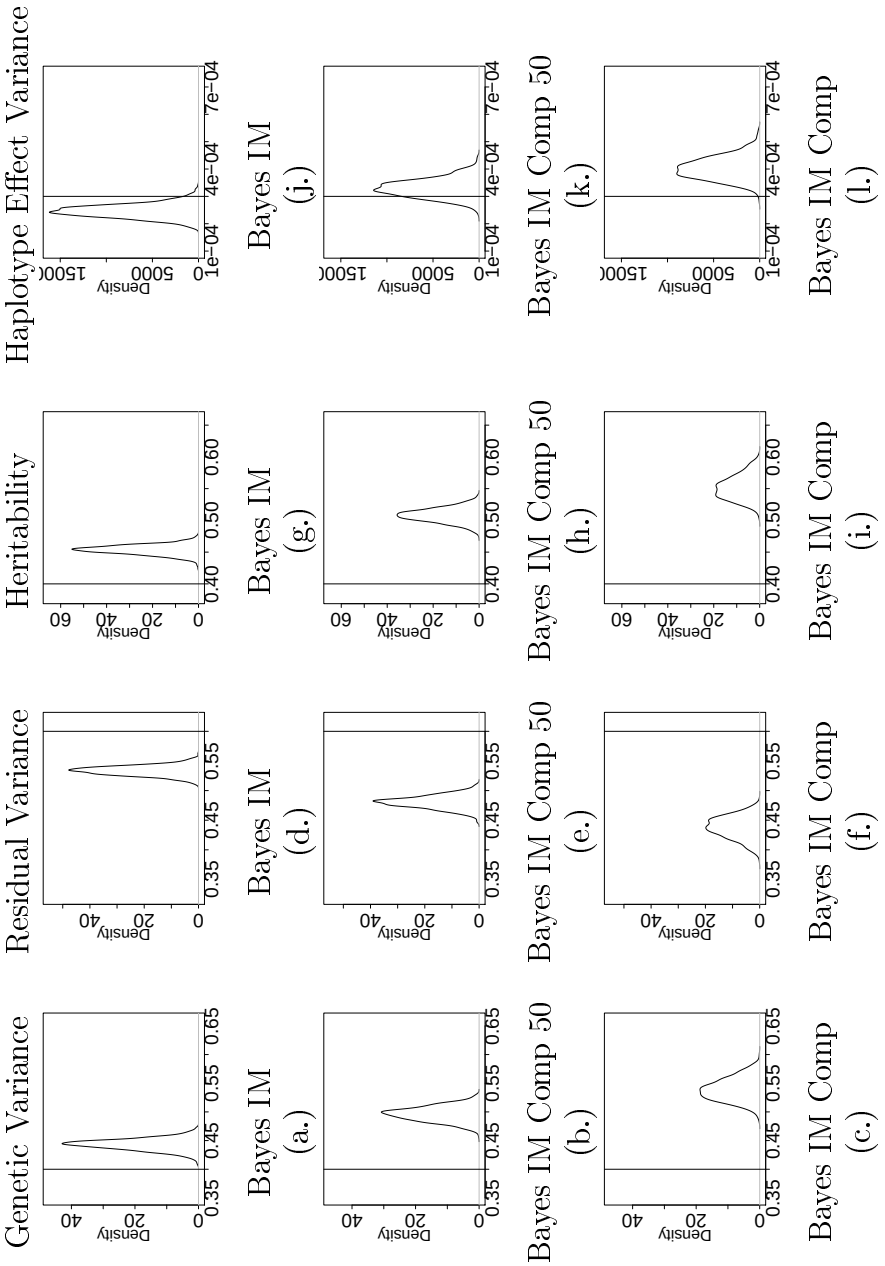
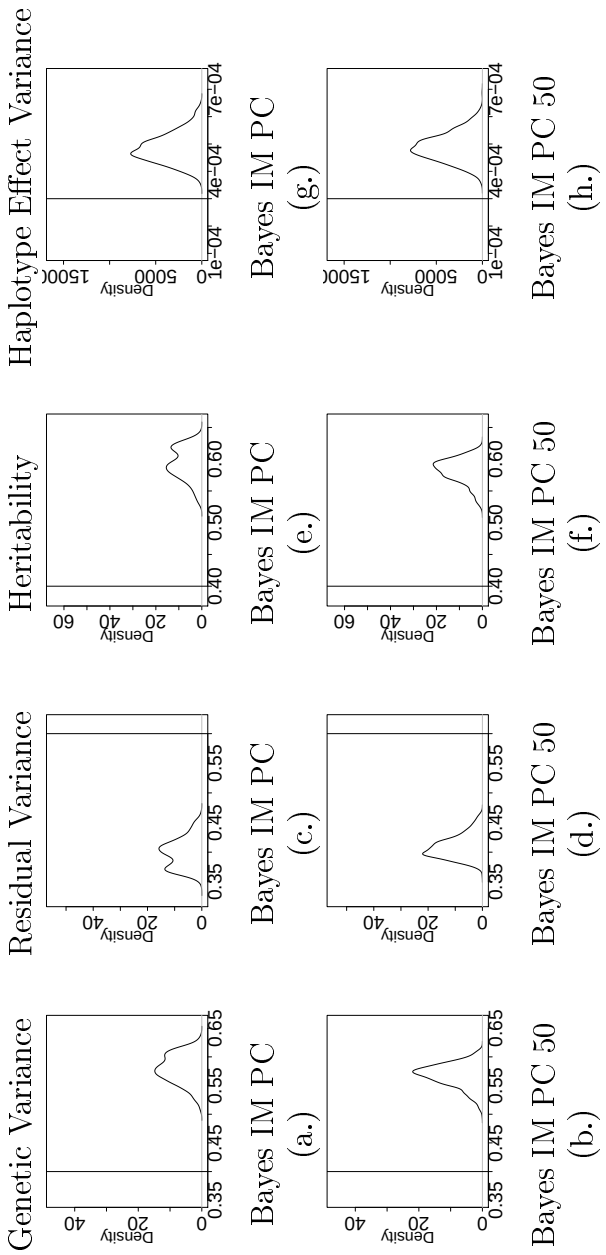
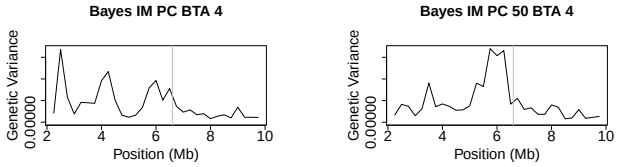


Figure B.2.3: Density Plots for the Variance Components in Bayes IM PC Models



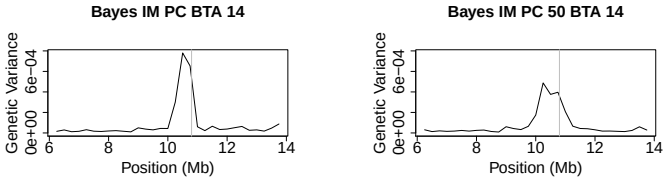
B.3 QTL Identification and Haplotype Effects

Figure B.3.1: QTL Identification for BTA 4 Between 2 and 10 MB



(a.) (b.)

Figure B.3.2: QTL identification for BTA 14 between 6 and 14 MB



(a.) (b.)

APPENDIX C

ADDITIONAL RESULTS FOR THE REPRODUCTIVE LONGEVITY DATA SET

C.1 Posterior Distributions

Figure C.1.1: Posterior Distribution of Parameters for the Bayes B and C Models in the Reproductive Longevity Data Set

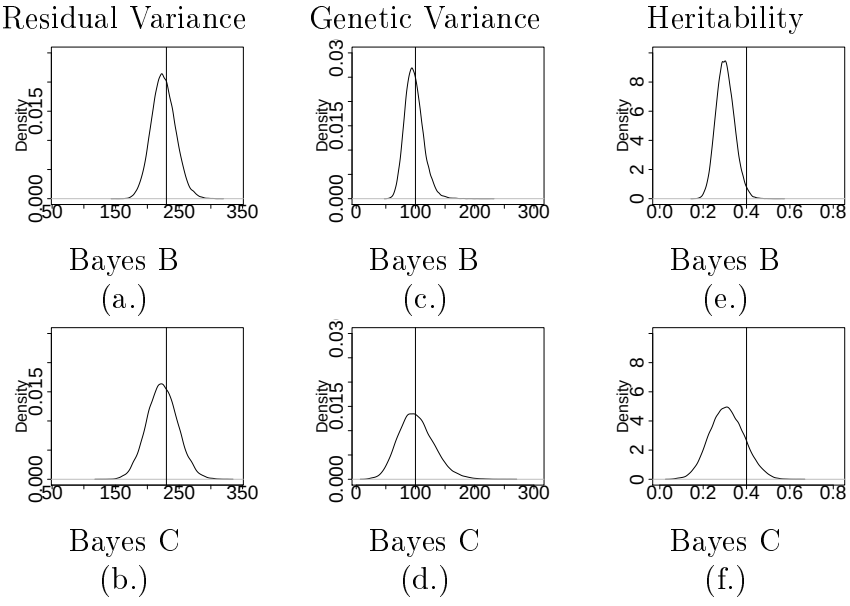


Figure C.1.2: Posterior Distribution of Parameters for the Bayes IM Models in the Reproductive Longevity Data Set

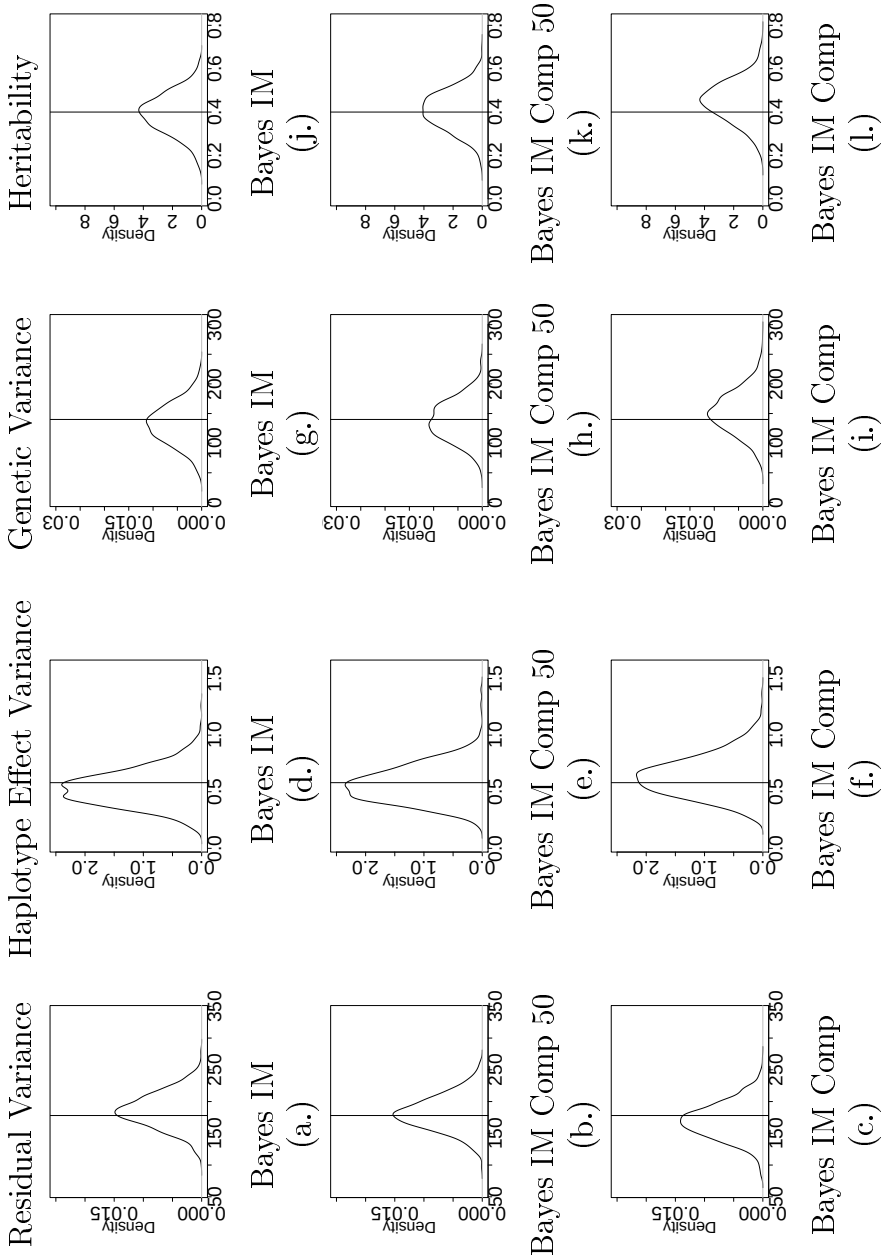
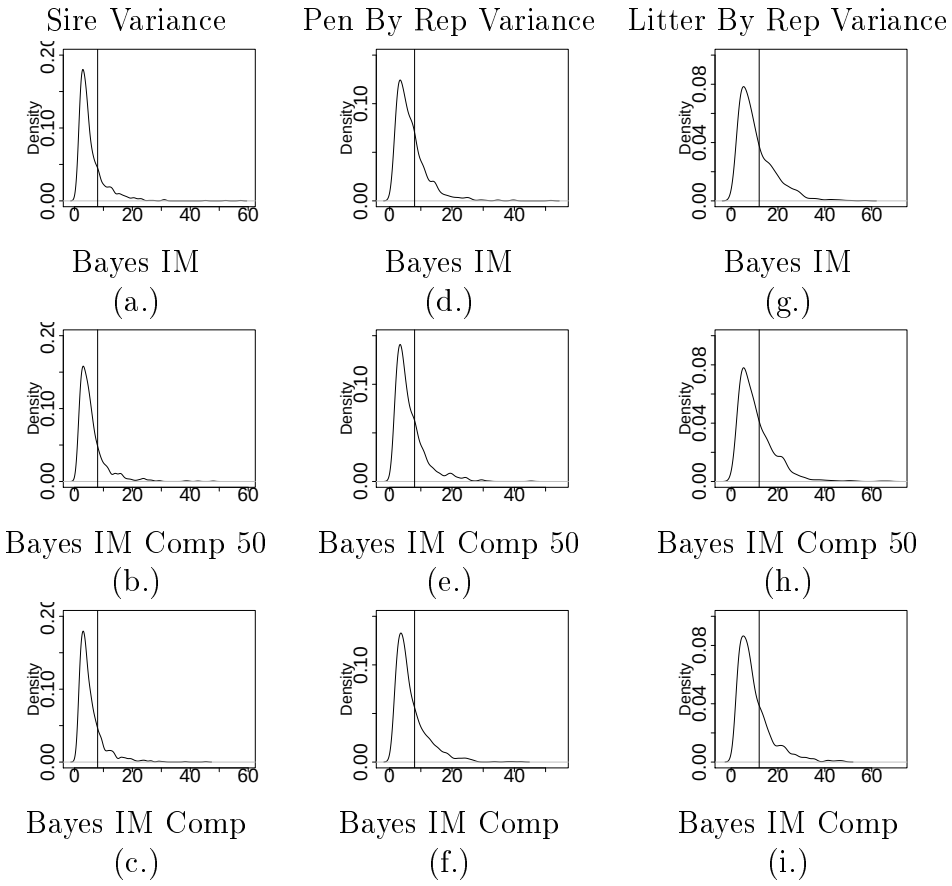


Figure C.1.3: Posterior Distributions for Random Effects in NIL



C.2 QTL Identification and Haplotype effects

Table C.2.1: Top Windows for Bayes IM Models

Rank	Bayes IM			
	SSC	Start	End	GenVar
1	**7**	**117,900,000**	**118,800,000**	**0.21828**
2	14	92,550,000	93,450,000	0.18144
3	6	149,550,000	150,450,000	0.177067
4	12	900,000	1,800,000	0.16986
5	14	91,650,000	92,550,000	0.169602
6	6	150,450,000	151,350,000	0.169412
7	5	55,650,000	56,550,000	0.167168
8	5	56,550,000	57,450,000	0.164331
9	14	93,450,000	94,350,000	0.160192
10	**6**	**87,450,000**	**88,350,000**	**0.160063**

Rank	Bayes IM PC 50			
	SSC	Start	End	GenVar
1	**7**	**117,900,000**	**118,800,000**	**0.225118**
2	12	1,800,000	2,700,000	0.210913
3	**2**	**12,600,000**	**13,500,000**	**0.20001**
4	5	53,850,000	54,750,000	0.191019
5	14	88,950,000	89,850,000	0.189228
6	3	132,150,000	133,050,000	0.186363
7	**2**	**13,500,000**	**14,400,000**	**0.183272**
8	6	150,450,000	151,350,000	0.178461
9	3	131,250,000	132,150,000	0.175847
10	5	54,750,000	55,650,000	0.175468

Rank	Bayes IM PC			
	SSC	Start	End	GenVar
1	**2**	**13,500,000**	**14,400,000**	**0.36055**
2	**6**	**88,350,000**	**89,250,000**	**0.320961**
3	**2**	**12,600,000**	**13,500,000**	**0.246828**
4	**6**	**89,250,000**	**90,150,000**	**0.231518**
5	10	8,850,000	9,750,000	0.221879
6	**6**	**87,450,000**	**88,350,000**	**0.218702**
7	3	132,150,000	133,050,000	0.214654
8	**2**	**14,400,000**	**15,300,000**	**0.212291**
9	3	133,050,000	133,950,000	0.206281
10	6	150,450,000	151,350,000	0.192561

a. Bold represents the identification of a QTL on SSC 2 between 12-15 MB, SSC 6 between 87-89 MB, or SSC 7 between 118-120 MB.

Table C.2.2: Top Windows for Bayes B and Bayes C Models

Rank	Bayes B			
	SSC	Start	End	% GenVar
1	**7**	**118,001,966**	**118,979,653**	**1.06**
2	**6**	**88,024,202**	**88,964,845**	**0.91**
3	5	4,092,291	4,916,323	0.78
4	9	113,051,200	113,942,105	0.46
5	6	101,036,334	101,986,509	0.42
6	6	29,076,106	29,958,963	0.39
7	12	2,065,310	2,979,580	0.39
8	14	93,096,651	93,968,851	0.33
9	6	35,011,158	35,967,061	0.31
10	14	50,012,997	50,992,818	0.31

Rank	Bayes C			
	SSC	Start	End	% Gen Var
1	6	35,011,158	35,967,061	0.30
2	**7**	**118,001,966**	**118,979,653**	**0.29**
3	5	4,092,291	4,916,323	0.29
4	14	50,012,997	50,992,818	0.24
5	7	40,029,052	40,992,238	0.24
6	**2**	**12,047,540**	**12,999,985**	**0.23**
7	9	11,012,156	11,997,728	0.20
8	6	29,076,106	29,958,963	0.19
9	**6**	**88,024,202**	**88,964,845**	**0.19**
10	12	27,013,251	27,972,588	0.18

a. Percent of Genetic Variance is reported here as this is the value provided by the GenSel software.

b. Bold represents the identification of a QTL on SSC 2 between 12-15 MB, SSC 6 between 87-89 MB, or SSC 7 between 118-120 MB.

Table C.2.3: Probability of Individual Cluster Membership on SSC 2

Cluster	Bayes IM	Bayes IM PC 50	Bayes IM PC
		NIL	NIL
1	0.0681	0.1038	0.0689
2	**0.0004**	0.0690	**0.0053**
3	0.1445	**0.0052**	0.1777
4	**0.0008**	**0.0049**	**0.0053**
		L2	
5	**0.0030**	**0.0063**	**0.0055**
6	0.0692	0.1605	0.1065
7	0.0801	0.0428	0.0633
8	0.0809	0.1291	0.0675
		Common	L2
9	0.0774	**0.0048**	**0.0049**
10	0.0888	**0.0048**	0.1001
11	0.1005	**0.0048**	**0.0058**
12	0.1862	0.0730	0.0311
13	0.0166	0.0768	0.0657
14	0.0822	0.1375	0.1436
15	**0.0007**	**0.0048**	**0.0053**
16	**0.0004**	0.1721	0.1434

a. Bold indicates where the cluster membership probability is less than 0.01.

Table C.2.4: Probability of Individual Cluster Membership on SSC 6

Cluster	Bayes IM	Bayes IM PC 50	Bayes IM PC
		NIL	NIL
1	**0.0002**	0.0653	0.3646
2	0.0439	**0.0048**	**0.0046**
3	0.2030	**0.0048**	0.0878
4	**0.0002**	**0.0048**	**0.0046**
		L2	
5	0.0592	0.0133	**0.0072**
6	**0.0002**	**0.0047**	0.0221
7	**0.0002**	0.0522	**0.0047**
8	**0.0002**	**0.0047**	**0.0046**
		Common	L2
9	**0.0002**	0.1466	**0.0076**
10	0.0249	0.1171	0.0518
11	0.0814	0.2218	**0.0048**
12	0.3185	0.1021	0.2369
13	0.2245	0.1569	0.1043
14	**0.0052**	0.0912	**0.0047**
15	**0.0002**	**0.0048**	0.0323
16	0.0380	**0.0048**	0.0576

a. Bold indicates where the cluster membership probability is less than 0.01.

Table C.2.5: Probability of Individual Cluster Membership on SSC 7

Cluster	Bayes IM	Bayes IM PC 50	Bayes IM PC
		NIL	NIL
1	0.0558	0.1066	0.0986
2	**0.0001**	0.2476	**0.0047**
3	**0.0001**	**0.0048**	0.0911
4	**0.0021**	**0.0048**	0.0252
		L2	
5	0.1633	0.0207	**0.0047**
6	0.0364	0.0681	0.0361
7	0.0538	0.0951	**0.0048**
8	0.1436	0.1226	0.2350
		Common	L2
9	0.0359	0.0826	0.0445
10	0.1443	0.0547	**0.0052**
11	0.0853	**0.0047**	0.0101
12	0.0738	**0.0047**	0.1526
13	0.0310	0.0318	0.0622
14	0.0383	0.1306	0.0541
15	0.0669	0.0161	0.1665
16	0.0691	**0.0047**	**0.0048**

a. Bold indicates where the cluster membership probability is less than 0.01.